ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές Μηχανικής Μάθησης για την Ανίχνευση Ομοιότητας/Ανομοιότητας - Εφαρμογές Εντοπισμού Κοινοτήτων σε Κοινωνικά Δίκτυα και Ελέγχου Λογοκλοπής σε Κείμενα

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

της

ΕΛΕΝΗΣ Κ. ΒΑΘΗ

Αθήνα, Δεκέμβριος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές Μηχανικής Μάθησης για την Ανίχνευση Ομοιότητας/Ανομοιότητας - Εφαρμογές Εντοπισμού Κοινοτήτων σε Κοινωνικά Δίκτυα και Ελέγχου Λογοκλοπής σε Κείμενα

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

της

ΕΛΕΝΗΣ Κ. ΒΑΘΗ

Συμβουλευτική Επιτροπή: Γεώργιος - Ανδρέας Σταφυλοπάτης
Γεώργιος Στάμου
Παναγιώτης Τσανάκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 23η Δεκεμβρίου 2019.

.....	.....	.....
Γ.-Α. Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γ. Στάμου Αν. Καθηγητής Ε.Μ.Π.	Π. Τσανάκας Καθηγητής Ε.Μ.Π.
.....	.....	.....
Γ. Μέντζας Καθηγητής Ε.Μ.Π.	Σ. Παπαβασιλείου Καθηγητής Ε.Μ.Π.	Δ. Τσουκαλάς Καθηγητής Ε.Μ.Π.

.....
M. Βαζιργιάννης
Καθηγητής Ο.Π.Α. &
École Polytechnique, Paris

Αθήνα, Δεκέμβριος 2019

.....

ΕΛΕΝΗ Κ. ΒΑΘΗ
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια, η πληροφορία που διακινείται ηλεκτρονικά έχει αυξηθεί σε μεγάλο βαθμό, γεγονός που επιβάλλει τη δημιουργία και χρήση νέων συστημάτων, ικανών να διαχειρίζονται μεγάλο όγκο πληροφορίας. Η Μηχανική Μάθηση και η Εξόρυξη Δεδομένων είναι δύο πεδία μελέτης, τα οποία επιτρέπουν την ανάλυση και ταξινόμηση πληροφορίας. Οι αλγόριθμοι μηχανικής μάθησης «μαθαίνουν» από τα ίδια τα δεδομένα, ανακαλύπτοντας μοτίβα, χωρίς τη χρήση ρητών οδηγιών. Στο πλαίσιο της διατριβής, μελετήθηκαν και υλοποιήθηκαν τρεις ξεχωριστές, αλλά συναφείς προσεγγίσεις, για την Ανίχνευση Κοινοτήτων και την Εγγενή Ανίχνευση Λογοκλοπής, οι οποίες κάνουν χρήση τεχνικών μηχανικής μάθησης.

Η Ανίχνευση Κοινοτήτων, ή αλλιώς ομαδοποίηση γράφου, είναι ένα από τα πιο δημοφιλή θέματα της σύγχρονης επιστήμης δικτύων, που επιχειρεί να λύσει το πρόβλημα του εντοπισμού της κοινοτικής δομής σε δίκτυα. Τα περισσότερα δίκτυα εμφανίζουν κοινοτική δομή, δηλαδή οι κορυφές τους είναι οργανωμένες σε ομάδες, που ονομάζονται κοινότητες, ομάδες ή συστάδες. Η ανίχνευση κοινοτήτων δεν είναι ένα σαφώς ορισμένο πρόβλημα, καθώς δεν υπάρχει ένας αυστηρός και καθολικά αποδεκτός ορισμός για το τι είναι κοινότητα. Ο ορισμός αλλάζει ανάλογα με την εφαρμογή, δηλαδή με το ερευνητικό ερώτημα που καλούμαστε κάθε φορά να απαντήσουμε ή το συγκεκριμένο σύστημα το οποίο βρίσκεται υπό μελέτη.

Στο πλαίσιο της διατριβής, μελετήθηκε το πρόβλημα της ανίχνευσης κοινοτήτων στα κοινωνικά δίκτυα και προτάθηκε μια μεθοδολογία για τον εντοπισμό όμοιων χρηστών στο Twitter. Οι κοινότητες ορίζονται ως ομάδες χρηστών με μεγαλύτερη πυκνότητα συνδέσεων μεταξύ τους παρά με το υπόλοιπο δίκτυο, που αλληλεπιδρούν ο ένας με τον άλλο και έχουν κοινά ενδιαφέροντα. Επομένως, η συγκεκριμένη μεθοδολογία δεν βασίζεται μόνο στην τοπολογία του δικτύου για να ομαδοποιήσει τους χρήστες σε κοινότητες, αλλά λαμβάνει επιπλέον υπ΄ όψιν το κείμενο που μοιράζονται οι χρήστες και τις αλληλεπιδράσεις τους. Αρχικά, ορίζονται έξι διαφορετικές μετρικές ομοιότητας, με βάση όλα τα χαρακτηριστικά στοιχεία του Twitter που παρέχουν πληροφορία για τις αλληλεπιδράσεις των χρηστών. Οι μετρικές αυτές συνδυάζονται, και ο συνδυασμός

τους χρησιμοποιείται για την ομαδοποίηση των χρηστών σε κοινότητες. Επίσης, παρουσιάζεται μια νέα μέθοδος που εξάγει τα θέματα που συζητούνται σε κάθε κοινότητα, με στόχο να εντοπιστούν τα ενδιαφέροντα των χρηστών. Ακόμα, προτείνεται μια μέθοδος αφαίρεσης των θεμάτων που δεν παρουσιάζουν ενδιαφέρον και περιγράφεται μια διαδικασία για την αυτόματη παραγωγή επισημάνσεων για κάθε θέμα.

Σε δεύτερη φάση, μελετάται η ενσωμάτωση γράφου και η εξαγωγή διανυσματικών παραστάσεων κόμβων. Οι μέθοδοι ενσωμάτωσης γράφου έχουν προταθεί ως εναλλακτική στις παραδοσιακές τεχνικές εξόρυξης γράφων. Στόχος τους είναι η μετατροπή ενός γράφου σε μια αναπαράσταση χαμηλών διαστάσεων, όπου κάθε κόμβος αντιστοιχεί σε ένα διάνυσμα χαμηλών διαστάσεων. Αυτά τα διανύσματα, που ονομάζονται, επίσης, διανυσματικές παραστάσεις κόμβων, μπορούν στη συνέχεια να δοθούν ως είσοδοι σε οποιονδήποτε αλγόριθμο επιβλεπόμενης μάθησης, μετατρέποντας, έτσι, το αρχικό πρόβλημα σε ένα ήδη γνωστό. Επομένως, οι μέθοδοι αυτές είναι χρήσιμες σε μια πληθώρα εφαρμογών του πραγματικού κόσμου, όπως είναι η ταξινόμηση κόμβων, η ανίχνευση κοινοτήτων, η πρόβλεψη συνδέσμου και η οπτικοποίηση δικτύων. Στα πλαίσια αυτά, προτείνεται η δεύτερη προσέγγιση της διατριβής, η οποία, σε αντίθεση με προηγούμενες προσεγγίσεις, οι οποίες λαμβάνουν υπ΄ όψιν μόνο τις ακμές ενός γράφου κατά την εξερεύνηση του μέσω τυχαίων περιπάτων, λαμβάνει επίσης υπ΄ όψιν τις ομοιότητες μεταξύ των κόμβων.

Η λογοκλοπή είναι η πράξη της αντιγραφής ή της μίμησης του έργου κάποιου άλλου και η παρουσίασή του ως πρωτότυπη, χωρίς όμως την κατάλληλη αναφορά ή παραπομπή. Η ανίχνευση λογοκλοπής σε έγγραφα κειμένου χωρίζεται σε δύο κύριες κατηγορίες, τις εξωγενείς και τις εγγενείς μεθόδους. Οι εξωγενείς μέθοδοι συγκρίνουν μια συλλογή εγγράφων, η οποία αποτελεί πιθανή πηγή προέλευσης των αντιγραμμένων αποσπασμάτων, και ένα σύνολο ύποπτων εγγράφων, ενώ οι εγγενείς μέθοδοι προσδιορίζουν ποια από τα αποσπάσματα του εγγράφου υπό διερεύνηση είναι αντιγραμμένα, παρατηρώντας τις διαφοροποιήσεις στον τρόπο γραφής μέσα στο ίδιο το κείμενο. Η κεντρική ιδέα στην οποία βασίζεται η εγγενής ανίχνευση λογοκλοπής είναι ότι κάθε συγγραφέας έχει το δικό του προσωπικό και μοναδικό στυλ γραφής, το οποίο μπορεί να ανιχνευθεί και να ποσοτικοποιηθεί χρησιμοποιώντας στυλιστικές ή/και σημασιολογικές τεχνικές.

Με βάση τα παραπάνω, παρουσιάζεται μια προσέγγιση εγγενούς ανίχνευσης λογοκλοπής για έγγραφα κειμένου. Αρχικά, προτείνεται μια σειρά νέων χαρακτηριστικών, τα οποία επιτρέπουν την ποσοτικοποίηση του τρόπου γραφής για κάθε απόσπασμα κειμένου. Τα χαρακτηριστικά αυτά συνδυάζονται με μια σειρά μοντέλων επιβλεπόμενης μάθησης, που εκπαιδεύονται να ταξινομούν τα αποσπάσματα ανάλογα με το αν έχουν προκύψει από λογοκλοπή ή όχι. Τέλος, μελετάται το πρόβλημα των μη ισορροπημένων δεδομένων, το οποίο αποτελεί

μία κρίσιμη παράμετρο του προβλήματος. Για το λόγο αυτό, εξετάζεται το κατά πόσον οι τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας βελτιώνουν τα αποτελέσματα του συστήματος.

Οι προτεινόμενες προσεγγίσεις αξιολογήθηκαν σε δημόσια διαθέσιμα σύνολα δεδομένων. Λόγω της φύσης του προβλήματος της ανίχνευσης κοινοτήτων, τα αποτελέσματα της πρώτης μεθοδολογίας δεν αξιολογούνται σε σύγκριση με κάποιον ήδη υπάρχοντα αλγόριθμο. Αντίθετα, η δεύτερη και η τρίτη προσέγγιση συγκρίθηκαν με τους state-of-the-art αλγορίθμους στο εκάστοτε πεδίο έρευνας. Τα αποτελέσματα των πειραμάτων αποδεικνύουν την ικανοποιητική συμπεριφορά των προτεινόμενων μεθοδολογιών, οι οποίες σε πολλές περιπτώσεις υπερτερούν σε σχέση με τους αλγορίθμους με τους οποίους συγκρίνονται.

Λέξεις-κλειδιά: Μηχανική Μάθηση, Ανίχνευση Κοινοτήτων, Ανάλυση Κοινωνικών Δικτύων, Μοντελοποίηση Θεμάτων, Διανυσματικές Παραστάσεις Κόμβων, Μάθηση Χαρακτηριστικών, Εγγενής Ανίχνευση Λογοκλοπής, Μη Ισορροπημένα Δεδομένα

ABSTRACT

In recent years, the amount of information transmitted online has greatly increased, which dictates the creation and usage of new systems, capable of handling large volumes of information. Machine Learning and Data Mining are two fields of study, which facilitate the analysis and classification of information. Machine learning algorithms "learn" directly from data, by discovering meaningful patterns, without the use of explicit instructions. In this thesis, three separate, but related, approaches for Community Detection and Intrinsic Plagiarism Detection, which utilize machine learning techniques, were studied and implemented.

Community Detection, or graph clustering, is one of the most popular topics in modern science of networks, which aims to solve the problem of identifying the community structure in networks. Most networks display community structure, i.e. the vertices are organized in groups, called communities, groups or clusters. Community detection is not a well defined problem, as there is no strict and universally accepted definition of community. The definition often depends on the application, the research question at hand or on the specific system under study.

In this thesis, the problem of community detection in social networks was studied and a methodology for identifying similar users on Twitter was proposed. Communities are defined as groups of users that are more densely connected to each other than to the rest of the network, interact more between them and share common interests. Therefore, this methodology does not solely depend on the network topology in order to group the users into communities, but also takes into account the text that users share, as well as their interactions. Initially, six different similarity metrics are defined, which are based on Twitter attributes that provide information regarding the interactions between users. These metrics are combined, and their combination is used for the clustering of users into communities. Additionally, a new method for extracting the topics discussed in each community is presented, which helps identify the users' interests. Also, a method for the elimination

of the trivial topics is proposed, and a process for automatically generating labels for the topics is described.

Secondly, graph embedding and the extraction of node embeddings are being studied. Graph embedding methods have been proposed as an alternative to traditional graph mining techniques. The objective is to convert a graph into a low dimensional representation, where each node of the graph would be mapped to a low dimensional vector. These vectors, also called node embeddings or feature vectors, can then be presented as input to any supervised learning algorithm, thus simplifying the original problem. Therefore, these methods can be useful in a variety of real-world applications, such as node classification, community detection, link prediction and network visualization. Based on the above, the second approach is proposed, which, contrary to previous approaches, which only take into account the edges of the graph when exploring the graph through random walks, also considers the similarities between the nodes.

Plagiarism is the act of taking or closely imitating someone else's work and presenting it as original, without proper citation or acknowledgment. Plagiarism detection in text documents is divided into two major categories, extrinsic and intrinsic methods. Extrinsic methods detect the suspicious similarities between a collection of potential source documents and a set of suspicious documents, while in intrinsic methods the objective is to identify which of the passages of an investigated document are plagiarized by observing the variation of the writing style within the document. Intrinsic plagiarism detection is based on the idea that every author has its own personal and unique writing style, which can be detected and quantified using stylistic and/or semantic means.

Based on the above, an intrinsic plagiarism detection approach for text documents is presented. Initially, a set of novel stylistic features, which help quantify the author's writing style for the whole document and each suspicious passage, is introduced. These features are then combined with a number of supervised learning methods, in order to classify the passages into plagiarized or non-plagiarized. Finally, the unbalanced nature of the datasets is examined, which is considered a crucial parameter for this task. As a result, over-sampling and under-sampling techniques are used in order to examine whether the performance of the proposed system is improved.

The proposed approaches were evaluated on publicly available datasets. Due to the nature of the problem of community detection, the results of the first methodology are not evaluated in comparison to an already existing approach. On the contrary, the second and third approaches were compared to the state-of-the-art algorithms for each respective field of study. The experimental results demonstrate the satisfactory behavior of the pro-

posed methodologies, which in many cases outperform the algorithms they are compared to.

Keywords: Machine Learning, Community detection, Social Networks Analysis, Topic Modeling, Node Embeddings, Feature Learning, Intrinsic Plagiarism Detection, Imbalanced Data

EXTENDED ABSTRACT

In recent years, the amount of information transmitted online has greatly increased, which dictates the creation and usage of new systems, capable of handling large volumes of information. Machine Learning and Data Mining are two fields of study, which facilitate the analysis and classification of information. Machine learning algorithms "learn" directly from data, by discovering meaningful patterns, without the use of explicit instructions.

In this thesis, machine learning techniques were adopted for the study and implementation of methodologies for Community Detection and Intrinsic Plagiarism Detection. In particular, three separate, but related, approaches are proposed, which use the concept of similarity (proximity) or dissimilarity (distance) between data, aiming to identify communities on social networks, extract embeddings in graphs and detect plagiarism in texts.

Community Detection, or graph clustering, is one of the most popular topics in modern science of networks, which aims to solve the problem of identifying the community structure in networks. Most networks display community structure, i.e. the vertices are organized in groups, called communities, groups or clusters. Revealing the underlying communities can shed light into the structural properties of real-world networks and the way these networks function. Community detection is not a well defined problem, as there is no strict and universally accepted definition of community. The definition often depends on the application, the research question at hand or on the specific system under study.

As a result, there is no explicit way of evaluating the performance of the various algorithms, which makes the comparison between different approaches difficult. On the one hand, this slows down the progress in solving an already particularly difficult computational problem, on the other hand this ambiguity leaves a great deal of freedom in the different approaches proposed for the problem.

The first proposed approach addresses the problem of detecting communities on social networks, and more specifically, identifying similar users on

Twitter. The detection of communities in social networks can be achieved through two different kinds of approaches, the topology-based approaches or the topic-based approaches, as discussed in [26]. However, the author suggests that community detection should consider both the graph structure and textual information of the networks, since communities detected by the topology-based approaches tend to contain different topics within each community, while meaningful topology-based sub-communities exist inside each topic-based community.

Additionally, topology-based methods, while successfully managing to detect communities, do not provide any insight on what is the topic of discussion that bonds together the users of each community. Therefore, many different community detection approaches combine the topology of social connections and the topic features in order to extract meaningful topics of discussion between users [57, 88, 59].

In this work, communities are defined as groups of users that are more densely connected to each other than to the rest of the network, interact more between them and share common interests. Therefore, this methodology does not solely depend on the network topology in order to group the users into communities, but also takes into account the text that users share, as well as their interactions.

The proposed methodology consists of two steps. Initially, we define the concept of user similarity between Twitter users and compute the distance between each pair of user. The similarity between a pair of Twitter users is derived from the interactions recorded in their tweeting history. Therefore, similarity can be computed based on all elements that measure or describe each user's interaction with other users: the user's lists of friends and followers, the hashtags included in their tweets, their replies to other users and the users which are otherwise mentioned in their tweets.

Based on these Twitter attributes, which provide information regarding the interactions between users, we define six different similarity metrics: *following relationship similarity*, *common followers similarity*, *common friends similarity*, *hashtag similarity*, *reply similarity* and *user mention similarity*. In order to obtain the total user similarity, these measures must be combined. In our approach, we adopted a linear combination of the individual similarity measures.

The second step of community detection involves the clustering of the users for the formation of communities, by taking into account the similarity measures for each pair of users. It is well known that there is a wide range of clustering algorithms. Since the absolute positions of the data points are not available, the chosen algorithm must require the measures of similarity between pairs of data points as input. Affinity Propagation [32] is an algo-

rithm that identifies exemplars among data points and forms clusters of data points around these exemplars. It takes as input a collection of real-valued similarities between data points, and returns a number of clusters, which is not predefined.

Additionally, a new method for extracting the topics discussed in each community is presented, which helps identify the users' interests. The extraction of topics can be achieved by means of latent Dirichlet allocation (LDA) [9], a generative probabilistic model of a corpus, which is based on the idea that documents are represented as random mixtures over latent topics, where each topic is a probability distribution over words.

By using the LDA algorithm on the users' collection of documents we obtain a set of topics (and the distribution of words or keys per topic) and a corresponding topic distribution for each user's document. Our goal is to find the topic distribution for each cluster, and by extension the interests of the users. Therefore, we can consider that all users in a cluster are forming a document, aggregate the cluster's documents and finally get a cluster specific document collection.

By calculating the topic distribution of this cluster collection, while keeping the same topics as with the users' collection, we obtain the topic distribution for each cluster. In a similar manner, we can compute the total topic distribution, by aggregating all documents. At this point, we can measure how much the discussion in a community deviates from the general discussion and focuses on specific subjects, by computing the distance between each cluster's topic distribution and the topic distribution for the whole collection.

From a mining perspective, the topics extracted using LDA are not of equal importance. Some topics consist of general, everyday words, while others represent common interests for the majority of the examined users. Therefore, we developed a novel method for eliminating trivial topics. The method consists of two different approaches, meaning that for each topic two distinct measures are computed, which are then combined in order to rank the topics from the most to the least interesting. The first measure is based on the idea that interesting topics are usually discussed in a small number of communities. Therefore, for each cluster we normalize the topic distributions and calculate each topic's in-cluster percentage.

When executing the LDA algorithm, for each topic we obtain a list consisting of the top m words. The significance of the topic can be inferred by examining how common these words are. This can be achieved by calculating the frequency of appearance of each word in a corpus of the English language. In this work, we utilized the latest Wikipedia dump in order to retrieve the frequency for each word. As a result, we computed the mean word frequency for the words belonging to the topic list, which constitutes

the second proposed measure. It is expected that trivial topics would have higher mean word frequency values compared to the interesting ones. In order to combine the two measures and rank the topics, we simply calculate the arithmetic mean of the two measures. We will refer to this measure as the Composite Ranking Index.

As explained previously, each topic is represented by a list of keywords. Usually, in order to reveal the semantics of a topic, further processing is necessary. As a result, we propose a methodology to automatically generate labels using content retrieved from the English Wikipedia. We access the required data from Wikipedia via the MediaWiki API, a web service that provides convenient access to wiki features, data, and metadata. For each keyword of a topic, we search and retrieve the most relevant Wikipedia pages.

In order to determine which of the retrieved pages is the most representative of the keyword, we compute the number of appearances of all keywords of the topic in the text of the page. The page with the highest value is chosen. This process results in a unique page per keyword. Once again, the MediaWiki API is used to retrieve a list of categories for each one of the remaining pages. All categories are gathered and sorted based on the number of pages to which they correspond. The most common categories are chosen as labels for the topic. The final result is a small set of words or short phrases, instead of a long list of keywords.

The dataset utilized in this work was collected using the Twitter Searching API. Since our aim was to include users with common interests in the dataset, we selected the followers of @isocpp as our collection of users. @isocpp is the Twitter account of the ISO C++ standards committee, so it is expected that the users following this account would be interested to some extent and tweet about programming. This resulted in a set of 5077 users. For each of these users we crawled all of their published tweets and we retrieved a list of their followers' IDs and a list of their friends' IDs.

In order to validate the parameters of our model, the Calinski-Harabasz index [17] is utilized. This clustering quality criterion focuses on high intra-cluster similarity and low inter-cluster similarity, so it ensures that there are similar topic compositions for the users within a cluster and dissimilar topic compositions between different clusters. The criterion is used in order to find the optimal values for the weights of the six individual similarity metrics, as well as for the optimal number of topics.

Finally, a new metric is be defined, called *Local Twitter Community Interestingness*. High values of this metric signify large communities, with interesting (non-trivial) topics, which discuss more specific subjects, since they differentiate from the general discussion.

The second proposed approach involves studying graph embedding and

the extraction of node embeddings. Graph embedding methods have been proposed as an alternative to traditional graph mining techniques. The objective is to convert a graph into a low dimensional representation, while preserving the structure of the network, where each node of the graph would be mapped to a low dimensional vector. These vectors, also called node embeddings or feature vectors, can then be presented as input to any supervised learning algorithm, thus simplifying the original problem. Therefore, these methods can be useful in a variety of real-world applications, such as node classification, community detection, link prediction and network visualization [110, 18, 40, 35, 108].

A number of recent graph embedding methods utilize random walks to learn the node embeddings [73, 38, 29]. Random walks can be an effective tool when exploring networks. They are especially efficient when the graph is too large, or when only parts of the graph are known, since they only use local information of the network. These methods take advantage of the recent advancements in Natural Language Processing and unsupervised feature learning on documents, by treating random walks as the equivalent of sentences.

Many popular community detection algorithms are based on random walks [31, 18, 110]. The reason is that, when a graph has strong community structure, the nodes inside the communities are more densely connected to each other than to the rest of the network, forcing the random walks to spend more time visiting nodes that belong to the same community. Based on this known property of random walks, a novel, unsupervised approach to graph embedding is proposed. Contrary to previous approaches, which only take into account the edges of the graph when exploring the graph through random walks, the proposed methodology also considers the similarities between nodes.

The intuition behind the above manifests in a simple idea: instead of randomly choosing among the neighbours of a node, to force the random walk to move to nodes that are similar to the current node. This results in representations that not only include the information about the edges between the nodes in a vector space, but also encode the similarities between each pair of nodes with respect to some property. By adjusting the probability of visiting each node, and setting it to be proportionate to the similarity between two nodes, the probability of traversing edges that lie between communities is minimized. As a result, the proposed methodology is expected to have better performance in graphs with strong community structure.

Random walk based approaches commonly consist of two main components, a random walk generator and an update procedure [73, 38]. This methodology follows the same principle, with the exception that three dif-

ferent random walk generators are examined. Each generator implements a different exploration procedure, preserving different properties of the network. The generators take a graph G as input, and return a set of random walks.

The first generator, called U, uses uniform random walks to sample the graph. At each step of the walk, the generator chooses the next node to visit by sampling uniformly from the neighbours of the current node. Basically, this generator is identical to the one from the Deepwalk methodology. This generator aims to preserve the connections (the edges) between the nodes. The second generator, S_nbr, chooses the next node based on the similarity between the current node and each of its neighbours. The third generator, S_{any}, also uses non-uniform random walks, however the choice of the next node at each step is not limited to the set of neighbours of the current node, therefore it aims to preserve the similarities between the nodes.

The intuition behind the second and third generator is that, by forcing the random walk to move to nodes that are similar to the current node, the probabilities of traversing edges between communities are minimized, so each walk is constrained within the limits of a community. Especially, for the third generator, the restriction that a random walk can only move between neighbouring nodes is removed. This is based on many traditional clustering methods, which compute the similarity between each pair of vertices with respect to some property, without considering whether they are connected by an edge or not.

There are many local similarity methods which use neighbourhood-related information to compute the similarity between two nodes in the network. In this work, we experiment with five well-known metrics: (i) the *number of common neighbours*, (ii) the *Jaccard index*, (iii) the *euclidean distance*, (iv) the *cosine similarity*, and (v) the *Pearson correlation*.

In order to produce the embedding vectors from the random walks, the Skip-gram algorithm is employed to learn the latent representations of the nodes. This model is a two-layer neural network which produces a vector space for words when given a large corpus as input. Each word in the corpus corresponds to a vector in the space, while words that appear within the same context in the corpus are located close to one another in the vector space. In the case of networks, the random walks are treated as equivalent of sentences, and a sliding window chooses the nodes which appear within the same context. The Skip-gram model can be trained either with hierarchical softmax [66] or with negative sampling [65], with the latter being the one used in this study, as it results in faster training.

When coupled with the update procedure, each of the proposed generators results in a different set of features. As a last step to the methodology, the

results from the different exploration procedures can be combined. This can happen by concatenating, for each node, the embedding vectors which resulted from the use of different random walk generators. It is important to note that in all experiments, the embedding vector has the same size, in spite of any concatenation taking place. This is accomplished by adjusting the size of each of the vectors being concatenated, i.e. d/2 for each vector, when two exploration procedures are combined; d/3 when all three exploration procedures are combined.

The proposed methodology is evaluated on a wide range of artificial and real-world networks against a couple of state-of-the-art graph embedding techniques. As a first step, the proposed methodology is evaluated on artificial networks generated by the Lancichinetti-Fortunato-Radicchi (LFR) benchmark [56]. The LFR algorithm produces artificial networks which have a priori known communities, therefore they can be used to compare different community detection methods. The second part of the evaluation involves real-world datasets. Specifically, the proposed methodology is evaluated on three datasets, *BlogCatalog* [100], *Protein-Protein Interactions (PPI)* [15] and *Wikipedia* [62]. Each of these networks were selected in [38] because they exhibit different mixes of homophilic and role equivalences.

The proposed model is evaluated on a multi-label classification task. In order to compare it to the other methods, the exact same experimental procedure is used as in [73]. The data is split into a training and test set by randomly sampling a portion of the nodes to use as training set, while the rest are used as test set. This process is repeated 10 times. The training data are then given as input to a one-vs-rest logistic regression classifier extended to return the most probable labels. The results are evaluated using the Macro-F1 and Micro-F1 scores. The performance of the proposed models are evaluated against Deepwalk and node2vec. In order to ensure a fair comparison, an equal number of training samples is generated for all methods, while the optimization is run for a single epoch for all models.

The results show that this methodology performs better than the state-of-the-art algorithms, when applied on networks where the labels reveal the community structure of the network. In all other cases, the results vary, depending on the chosen strategy. However, certain strategies outperform the state-of-the-art algorithms, even when compared to semi-supervised approaches.

The best performing strategy overall for embedding the graph seems to be the combination of the first and third generators (namely $U + S_{any}$), meaning that the best results are achieved when the information concerning the connections and the similarities is encoded in different features (different columns in the final embedding matrix), so that the classifier using these

features can determine which apply better to the problem at hand. This confirms that the similarities between the nodes are equally important to the connections between them, and the use of the information from both attributes can lead to better embeddings.

Finally, the performance of the system is examined for different similarity metrics. It is shown that the choice of the similarity metric does not affect significantly the results, as the system performs equally well regardless of the chosen metric.

The third proposed approach is an Intrinsic Plagiarism Detection system. Plagiarism is the act of taking or closely imitating someone else's work and presenting it as original, without proper citation or acknowledgment. Plagiarism detection in text documents is divided into two major categories, extrinsic and intrinsic methods, respectively. The difference between them is whether a reference collection of source documents is required. Consequently, extrinsic methods detect the suspicious similarities between a collection of potential source documents and a set of suspicious documents, while in intrinsic methods the objective is to identify which of the passages of an investigated document are plagiarized by observing the variation of the writing style within the document.

Intrinsic plagiarism detection is based on the idea that, not only every author has its own personal and unique writing style, but, by using stylistic and/or semantic means, this style can be detected and quantified. As a result, by analyzing a document and searching for passages that do not seem to fit the personal writing style of the author, it is also possible to detect potential plagiarism.

Based on the above, an intrinsic plagiarism detection approach for text documents is presented. The system consists of one pre-processing step, three major parts (text segmentation, style analysis - feature extraction, outliers detection) and a results' post-processing step. For the text segmentation we apply the sliding window method for two different configurations: fixed-value parameters and 3-scale parameters' values according to the size of the document under examination. For the style analysis part we extract 11 features, both stylistic and semantic. While 4 of these features are used by most existing systems, the rest of them are novel features, which are designed based on the idea of compression or on the word frequency class concept. For the outlier detection part we rely on machine learning. We experiment with three classifiers: Decision Trees, Support Vector Machines and Random Forests.

The intrinsic plagiarism detection task is by nature a task of unbalanced data: the plagiarized sections tend to be significantly less in number than the original ones. This fact is critical for a supervised machine learning

classification. We address the problem by applying balancing techniques. We use and compare the results of a number of different methods. Initially, we balance the training data using the SMOTE-borderline algorithm [12]. SMOTE-borderline constructs synthetic examples of the minority class using the values of the existing minority class's examples and its neighbors, as long as they also belong to the minority class.

Our second approach is to experiment with a method which combines over- and under-sampling methods, called SMOTE + ENN [5]. SMOTE + ENN is a balancing method resulting from the application of the SMOTE-borderline algorithm followed by the Edited Nearest Neighbors (ENN) [112] rule. ENN is an under-sampling technique which removes any example whose class label differs from the class of the majority of its nearest neighbors. As a result, it "edits" the dataset by removing samples which do not agree "enough" with their neighborhood, so it affects the noisy samples next to the boundaries of the classes, and, therefore, can eliminate class overlapping. The motivation behind this approach is that SMOTE can generate noisy samples by interpolating new points between marginal outliers and inliers, which can be solved by cleaning the resulted space obtained after over-sampling, using ENN.

Finally, we experiment with the first step of the SMOTE + ENN method by applying different rates of oversampling. Instead of resampling to full balance, we increase the size of the minority class to a proportion of the size of the majority class. The reasoning behind this choice is that the ratio of plagiarized to non plagiarized segments in our dataset can be as small as 1:20, which means that using an oversampling method to equalize the sizes of both classes can practically double the size of our dataset. By extension, this increases the likelihood of over-fitting for the minority class, while slowing down the training process.

The proposed system is evaluated on the datasets constructed for the PAN Webis competitions (years 2009, 2011 and 2016). PAN is a series of scientific events and shared tasks on digital text forensics. We experiment with all the parameters of the system discussed so far: for the text segmentation part we experiment both with fixed and varying sliding window length; for the machine learning part we experiment with different pairs of training data balancing methods and classifiers.

The results show that Support Vector Machines do not correspond well to unbalanced training data; the classifier is not capable of learning from the instances of the minority class. Moreover, Random Forest and Decision Trees give much poorer results when the training data is unbalanced. As for the balancing methods, they all contribute to better performance in any classifier. However, it seems that there is no obvious winner among them, as

each classifier behaves best for different balancing method.

The proposed system is compared to other systems on PAN'09 and PAN'11 datasets. In both datasets our system achieves the best results. In particular, in the 2009 corpus we achieve more than 10% higher F-score than the best participants, while in the 2011 corpus our system succeeds about 5% higher F-score. In most cases, even when a subset of the features is used, the system outperforms the systems it is compared to.

Ευχαριστίες

Η διατριβή αυτή εκπονήθηκε στο Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης, υπό την επίβλεψη του καθηγητή κ. Ανδρέα-Γεώργιου Σταφυλοπάτη, τον οποίο θα ήθελα να ευχαριστήσω θερμά για την εμπιστοσύνη που έδειξε στο πρόσωπό μου, καθώς και για την πολύτιμη βοήθεια και υποστήριξη που μου προσέφερε σε όλη τη διάρκεια του διδακτορικού. Ευχαριστίες οφείλω, επίσης, στα άλλα δύο μέλη της συμβουλευτικής επιτροπής της διδακτορικής μου διατριβής, στον Αναπληρωτή Καθηγητή Ε.Μ.Π. κ. Γεώργιο Στάμου και στον Καθηγητή Ε.Μ.Π. κ. Παναγιώτη Τσανάκα, για τις συμβουλές και τις υποδείξεις τους, καθώς και στους Καθηγητές Ε.Μ.Π. κ. Γρηγόριο Μέντζα, κ. Συμεών Παπαβασιλείου και κ. Δημήτριο Τσουκαλά, και στον κ. Μιχάλη Βαζιργιάννη, Καθηγητή Ο.Π.Α. & École Polytechnique, για την τιμή που μου έκαναν να είναι μέλη της επιτροπής αξιολόγησης της διατριβής.

Στη συνέχεια, θα ήθελα να ευχαριστήσω όλα τα μέλη του εργαστηρίου για την άριστη συνεργασία και τη βοήθεια κατά τη διάρκεια του διδακτορικού, τα μέλη Ε.ΔΙ.Π. Γεώργιο Σιόλα και Γεώργιο Αλεξανδρίδη, τους διδάκτορες Αριστείδη Λαναρίδη, Γεώργιο Στρατογιάννη και Αγγελική Βλαχοστεργίου, καθώς και τους υποψήφιους διδάκτορες Αθανάσιο Τάγαρη, Μαρία Σδράκα, Παναγιώτη Κουρή, Γεώργιο Ιωάννου, Αναστάσιο Παπαγιάννη και Αθανάσιο Τασάκο. Ακόμα, θα ήθελα να ευχαριστήσω την Ανδριάννα Πολυδούρη, η οποία με εμπιστεύτηκε με τη συνέχιση της ήδη εξαιρετικής δουλειάς της.

Τέλος, ένα μεγάλο ευχαριστώ οφείλω στην οικογένειά μου, στους γονείς μου, Ευτέρπη και Κωνσταντίνο, στα αδέρφια μου, Μαρία, Νίκο και Φήβη, και στον Άγγελο, για όλη τη συμπαράσταση και τη στήριξη κατά τη διάρκεια των σπουδών μου.

Κεφάλαιο 1

Εισαγωγή

Τα τελευταία χρόνια, ο ολοένα αυξανόμενος όγκος πληροφορίας που διακινείται ηλεκτρονικά, σε συνδυασμό με την αυξημένη δημοτικότητα των κοινωνικών δικτύων, έχει καταστήσει επιτακτική την ανάγκη για νέα συστήματα τα οποία διαχειρίζονται, αναλύουν και ταξινομούν όλη αυτή την πληροφορία. Η Μηχανική Μάθηση και η Εξόρυξη Γνώσης από Δεδομένα είναι δύο πεδία μελέτης, τα οποία επιτρέπουν την ανάλυση και ταξινόμηση μεγάλου όγκου πληροφορίας.

Στο πλαίσιο της παρούσας διατριβής, μελετήθηκαν και υλοποιήθηκαν μεθοδολογίες για την Ανίχνευση Κοινοτήτων και την Εγγενή Ανίχνευση Λογοκλοπής σε κείμενα, με τη χρήση τεχνικών μηχανικής μάθησης. Ειδικότερα, προτείνονται τρεις ξεχωριστές, αλλά συναφείς προσεγγίσεις, οι οποίες χρησιμοποιούν την έννοια της ομοιότητας (της εγγύτητας), ή της ανομοιότητας (της απόστασης) των δεδομένων μεταξύ τους, με στόχο την ανίχνευση κοινοτήτων σε κοινωνικά δίκτυα, την ενσωμάτωση γράφου και την εγγενή ανίχνευση λογοκλοπής σε κείμενα.

Η ανίχνευση κοινοτήτων αποτελεί ένα σημαντικό ζήτημα της σύγχρονης επιστήμης δικτύων, αφού τα περισσότερα πραγματικά συστήματα παρουσιάζουν κοινοτική δομή, δηλαδή οι κορυφές των γράφων που τα αναπαριστούν μπορούν να χωριστούν σε ομάδες, ώστε πολλές ακμές να ενώνουν τις κορυφές της ίδιας ομάδας και λιγότερες ακμές να ενώνουν κορυφές διαφορετικών ομάδων.

Η ανίχνευση των κοινοτήτων σε ένα γράφο είναι σημαντική για διαφορετικούς τομείς της επιστήμης, όπως είναι η βιολογία, η επιστήμη των υπολογιστών, οι κοινωνικές επιστήμες. Ο εντοπισμός και η ανίχνευση κοινοτήτων δεν είναι μόνο ύψιστης σημασίας από ερευνητικής σκοπιάς, αλλά έχουν άμεσες εφαρμογές, για παράδειγμα στα συστήματα συστάσεων (recommender systems), στην εύρεση χρηστών που έχουν παρόμοια ενδιαφέροντα και στη στοχευμένη προώθηση αγαθών (marketing).

Η ενσωμάτωση γράφου είναι μια τεχνική εξόρυξης γράφων που μετατρέπει

ένα γράφο σε μια αναπαράσταση χαμηλών διαστάσεων. Οι κόμβοι αναπαρίστανται από διανύσματα χαμηλών διαστάσεων, τα οποία ονομάζονται διανυσματικές παραστάσεις κόμβων, και τα οποία έχουν την ιδιότητα να «κωδικοποιούν» πληροφορία σχετικά με τη δομή του γράφου. Τα διανύσματα αυτά, μπορούν στη συνέχεια να αποτελέσουν την είσοδο για κάποια μέθοδο μηχανικής μάθησης, μετατρέποντας το πρόβλημα της εξόρυξης γράφου σε πρόβλημα ταξινόμησης ή ομαδοποίησης.

Η τρίτη ερευνητική περιοχή που μελετάται είναι η εγγενής ανίχνευση λογοκλοπής σε έγγραφα κειμένου. Λογοκλοπή ονομάζεται η οικειοποίηση του έργου, των ιδεών ή των λέξεων ενός τρίτου και η παρουσίαση αυτών ως πρωτότυπων, χωρίς κάποια αναφορά στην αρχική πηγή. Οι μέθοδοι ανίχνευσης λογοκλοπής σε έγγραφα κειμένου χωρίζονται στις εξωγενείς και τις εγγενείς μεθόδους, ανάλογα με το αν απαιτείται, ή όχι, μια εξωτερική πηγή για την ανίχνευση της λογοκλοπής. Επομένως, οι εγγενείς μέθοδοι ανίχνευσης της λογοκλοπής προσδιορίζουν ποια από τα αποσπάσματα ενός εγγράφου είναι αντιγραμμένα, παρατηρώντας τις διαφοροποιήσεις στον τρόπο γραφής μέσα στο ίδιο το κείμενο, χωρίς να απαιτείται κάποια εξωτερική συλλογή εγγράφων.

1.1 Προτεινόμενες προσεγγίσεις

Στο πλαίσιο της διατριβής, μελετήθηκαν και αναπτύχθηκαν τρεις μεθοδολογίες, οι οποίες αναλύουν και ταξινομούν πληροφορία με τη χρήση τεχνικών μηχανικής μάθησης και είναι βασισμένες στις ιδέες της ομοιότητας ή της ανομοιότητας.

Όσον αφορά το πρόβλημα της ανίχνευσης κοινοτήτων στα κοινωνικά δίκτυα, προτείνεται μια μεθοδολογία που εντοπίζει όμοιους χρήστες στο Twitter. Η συγκεκριμένη μεθοδολογία δεν βασίζεται μόνο στην τοπολογία του δικτύου για να ομαδοποιήσει τους χρήστες σε κοινότητες, αλλά, προκειμένου να ορίσει την ομοιότητα ανάμεσα στους χρήστες, λαμβάνει επιπλέον υπ΄ όψιν το κείμενο που μοιράζονται οι χρήστες και τις αλληλεπιδράσεις τους. Επίσης, προτείνεται μια νέα μέθοδος που εξάγει τα θέματα που συζητούνται σε κάθε κοινότητα, με στόχο να εντοπιστούν τα ενδιαφέροντα των χρηστών.

Η δεύτερη προσέγγιση αφορά την εξαγωγή διανυσματικών παραστάσεων κόμβων. Η εν λόγω μεθοδολογία χρησιμοποιεί τυχαίους περιπάτους για την εξερεύνηση του γράφου, και, σε αντίθεση με προηγούμενες προσεγγίσεις, λαμβάνει υπόψη τις ομοιότητες μεταξύ των κόμβων κατά την εξερεύνηση. Στη συνέχεια, γίνεται η εξαγωγή των αναπαραστάσεων των κόμβων με τη χρήση των πρόσφατων εξελίξεων στη μη επιβλεπόμενη εξαγωγή χαρακτηριστικών από έγγραφα, αντιμετωπίζοντας τους τυχαίους περιπάτους ως το ισοδύναμο προτάσεων.

Τέλος, παρουσιάζεται μια προσέγγιση εγγενούς ανίχνευσης λογοκλοπής για έγγραφα κειμένου. Η προτεινόμενη μεθοδολογία συνδυάζει μια σειρά χαρακτηριστικών, που επιτρέπουν την ποσοτικοποίηση του τρόπου γραφής για κάθε απόσπασμα κειμένου, με την εφαρμογή μιας σειράς μεθόδων επιβλεπόμενης μάθησης που ταξινομούν τα αποσπάσματα ανάλογα με το αν έχουν προκύψει από λογοκλοπή ή όχι. Ταυτόχρονα, μελετάται κατά πόσον οι τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας βελτιώνουν τα αποτελέσματα του συστήματος.

1.2 Συνεισφορά της Διατριβής

Η συνεισφορά της διατριβής μπορεί να συνοψιστεί στα παρακάτω σημεία:

1. Παρουσίαση μιας μεθοδολογίας για τον εντοπισμό όμοιων χρηστών στο Twitter και κατ΄ επέκταση την ομαδοποίηση αυτών των χρηστών σε κοινότητες, η οποία λαμβάνει υπ΄ όψιν διάφορα χαρακτηριστικά του εν λόγω κοινωνικού δικτύου, όπως είναι το κειμενικό περιεχόμενο που μοιράζονται οι χρήστες, οι σχέσεις μεταξύ των χρηστών και οι αλληλεπιδράσεις τους.

2. Μελέτη των ενδιαφερόντων των χρηστών, με μια νέα μέθοδο που βασίζεται στον αλγόριθμο Λανθάνουσας Ανάθεσης Dirichlet, η οποία εξάγει τα θέματα που συζητούνται σε κάθε κοινότητα και εξαλείφει εκείνα που αποτελούνται από καθημερινές λέξεις.

3. Εισαγωγή μιας μεθόδου για την αυτόματη δημιουργία επισημάνσεων για τα θέματα.

4. Παρουσίαση μιας μεθόδου εξαγωγής διανυσματικών παραστάσεων κόμβων, βασισμένη στους τυχαίους περίπατους. Σε αντίθεση με προηγούμενες προσεγγίσεις, η προτεινόμενη μεθοδολογία λαμβάνει υπ΄ όψιν τόσο τις ακμές ενός γράφου, όσο και τις ομοιότητες μεταξύ των κόμβων, κατά την εξερεύνηση του γράφου.

5. Παρουσίαση μιας προσέγγισης εγγενούς ανίχνευσης λογοκλοπής, βασισμένης σε μεθόδους μηχανικής μάθησης.

6. Μελέτη του πως επηρεάζει την απόδοση ενός συστήματος εγγενούς ανίχνευσης λογοκλοπής η χρήση διαφορετικών τεχνικών εξισορρόπησης των δεδομένων, εφ΄ όσον η ύπαρξη μη ισορροπημένων δεδομένων αποτελεί κρίσιμη παράμετρο του προβλήματος.

1.3 Δομή της Διατριβής

Η παρούσα διατριβή είναι διαρθρωμένη ως εξής. Στο Κεφάλαιο 2 γίνεται μια σύντομη εισαγωγή στη Μηχανική Μάθηση, καθώς και μια εκτενής ανάλυση των μοντέλων Μηχανικής μάθησης που χρησιμοποιήθηκαν στις προτεινόμενες μεθοδολογίες. Στο Κεφάλαιο 3 γίνεται μια εισαγωγή στην Επεξεργασία Φυσικής Γλώσσας, όπου παρουσιάζονται συνοπτικά τα βήματα προεπεξεργασίας κειμένου και οι πιο διαδεδομένοι τρόποι αναπαράστασης κειμένου.

Στο Κεφάλαιο 4, γίνεται μια εισαγωγή στην ανίχνευση κοινοτήτων, παρουσιάζονται οι πτυχές του εν λόγω ερευνητικού προβλήματος, ενώ αναφέρονται οι σημαντικότερες προσεγγίσεις για την εφαρμογή της ανίχνευσης κοινοτήτων στα κοινωνικά δίκτυα. Στη συνέχεια αναπτύσσεται η μεθοδολογία η οποία προτείνεται σε αυτή την εργασία για την ανίχνευση κοινοτήτων στο Twitter. Τα πειραματικά αποτελέσματα για τη μεθοδολογία παρουσιάζονται στο Κεφάλαιο 5.

Το θεωρητικό υπόβαθρο για τις διανυσματικές παραστάσεις κόμβων δίνεται στο Κεφάλαιο 6, όπου γίνεται και η περιγραφή της μεθοδολογίας εξαγωγής διανυσματικών παραστάσεων κόμβων, της οποίας τα πειραματικά αποτελέσματα δίνονται στο Κεφάλαιο 7.

Στο Κεφάλαιο 8 παρουσιάζεται το θεωρητικό υπόβαθρο για την εγγενή ανίχνευση λογοκλοπής και περιγράφεται το προτεινόμενο σύστημα. Η αξιολόγηση της απόδοσης του συστήματος και η σύγκρισή του με άλλα αντίστοιχα συστήματα γίνεται στο Κεφάλαιο 9.

Στο Κεφάλαιο 10 παρουσιάζονται τα συμπεράσματα και οι μελλοντικές κατευθύνσεις έρευνας.

Κεφάλαιο 2

Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας από τους ταχύτερα αναπτυσσόμενους τομείς της επιστήμης υπολογιστών, με εκτεταμένες εφαρμογές σε διάφορα πεδία. Ο όρος μηχανική μάθηση αναφέρεται στην αυτοματοποιημένη ανίχνευση σημαντικών μοτίβων σε δεδομένα. Τις τελευταίες δεκαετίες έχει γίνει ένα κοινό εργαλείο σχεδόν σε κάθε εργασία που απαιτεί εξαγωγή πληροφοριών από μεγάλα σύνολα δεδομένων.

Η μηχανική μάθηση θεωρείται υποσύνολο της τεχνητής νοημοσύνης, ενώ σχετίζεται άμεσα με τα πεδία της στατιστικής και της εξόρυξης γνώσης από δεδομένα (data mining). Η μηχανική μάθηση και η εξόρυξη γνώσης από δεδομένα παρουσιάζουν σημαντική επικάλυψη, αφού συχνά χρησιμοποιούν τις ίδιες μεθόδους, αλλά με διαφορετικό στόχο. Η μηχανική μάθηση επικεντρώνεται στη μάθηση από τα δεδομένα εκπαίδευσης με στόχο την πρόβλεψη, ενώ η εξόρυξη γνώσης από δεδομένα επικεντρώνεται στην εξαγωγή άγνωστων ιδιοτήτων των δεδομένων. Συχνά, μάλιστα, οι μέθοδοι εξόρυξης γνώσης από δεδομένα χρησιμοποιούνται ως ένα βήμα προεπεξεργασίας για τη βελτίωση της ακρίβειας των μοντέλων μηχανικής μάθησης.

2.1 Κατηγορίες αλγορίθμων μάθησης

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να χωριστούν σε κάποιες ευρείες κατηγορίες. Οι τρεις πιο σημαντικές κατηγορίες είναι η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning).

Στην *επιβλεπόμενη μάθηση*, χρησιμοποιούνται σύνολα δεδομένων τα οποία περιέχουν τόσο τις εισόδους, όσο και τις επιθυμητές εξόδους, και ο στόχος είναι η δημιουργία ενός μαθηματικού μοντέλου το οποίο αντιστοιχίζει τα ζεύγη εισόδων-εξόδων. Κάθε παράδειγμα εκπαίδευσης του συνόλου δεδομένων

αποτελείται από την είσοδο, η οποία συνήθως δίνεται με τη μορφή ενός διανύσματος, και την έξοδο, η οποία αποτελεί την επισήμανση (label) της κλάσης στην οποία ανήκει το συγκεκριμένο παράδειγμα. Μετά την εκπαίδευση, ο αλγόριθμος επιβλεπόμενης μάθησης έχει μάθει μια συνάρτηση, η οποία μπορεί να χρησιμοποιηθεί για την αντιστοίχιση νέων, άγνωστων παραδειγμάτων, ενώ στη βέλτιστη περίπτωση, ο αλγόριθμος μπορεί να συμπεράνει σωστά την επισήμανση της κλάσης των νέων αυτών παραδειγμάτων.

Η *ταξινόμηση* (classification) και η *παλινδρόμηση* (regression) είναι δύο τύποι επιβλεπόμενης μάθησης. Οι αλγόριθμοι ταξινόμησης (ή ταξινομητές) χρησιμοποιούνται όταν οι έξοδοι περιορίζονται σε συγκεκριμένο σύνολο τιμών, επομένως οι μεταβλητές εξόδου είναι κατηγορικές (categorical) ή διακριτές. Οι αλγόριθμοι παλινδρόμησης έχουν συνεχείς εξόδους, άρα οι έξοδοι μπορούν να πάρουν οποιαδήποτε τιμή μέσα σε ένα διάστημα.

Στη *μη επιβλεπόμενη μάθηση*, δημιουργείται ένα μαθηματικό μοντέλο, το οποίο αποκαλύπτει άγνωστα μοτίβα στο σύνολο δεδομένων, χωρίς τη χρήση επισημάνσεων. Οι αλγόριθμοι μη επιβλεπόμενης μάθησης χρησιμοποιούνται για την εύρεση δομής στα δεδομένα, ενώ μπορούν να ομαδοποιούν τα δεδομένα σε κατηγορίες. Η μη επιβλεπόμενη μάθηση περιλαμβάνει την ομαδοποίηση ή συσταδοποίηση (clustering) δεδομένων, τη μάθηση χαρακτηριστικών (feature learning) και τη μείωση διαστάσεων (dimensionality reduction).

Η *ημι-επιβλεπόμενη μάθηση* είναι ένα υβρίδιο επιβλεπόμενων και μη επιβλεπόμενων τεχνικών, όπου οι αλγόριθμοι εκπαιδεύονται σε ελλιπή δεδομένα, όπου μόνο ένα ποσοστό των παραδειγμάτων εισόδου έχει επισημάνσεις.

Τέλος, η *ενισχυτική μάθηση* αποτελείται από αλγόριθμους που λαμβάνουν ανατροφοδότηση με τη μορφή θετικής ή αρνητικής ενίσχυσης σε ένα δυναμικό περιβάλλον, και δρουν με στόχο τη μεγιστοποίηση κάποιας έννοιας αθροιστικής ανταμοιβής.

2.2 Μοντέλα επιβλεπόμενης μάθησης

Σε αυτή την ενότητα, γίνεται εκτενής ανάλυση των μοντέλων μηχανικής μάθησης που χρησιμοποιήθηκαν στην παρούσα διατριβή, με στόχο την κατανόηση των προτεινόμενων προσεγγίσεων. Ειδικότερα, τα μοντέλα επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν είναι τα Πολυστρωματικά Perceptron, τα Δένδρα Αποφάσεων, τα Τυχαία Δάση και οι Μηχανές Διανυσμάτων Υποστήριξης, ενώ για τη μη επιβλεπόμενη μάθηση χρησιμοποιήθηκε ένας αλγόριθμος συσταδοποίησης, η Διάδοση Συνάφειας.

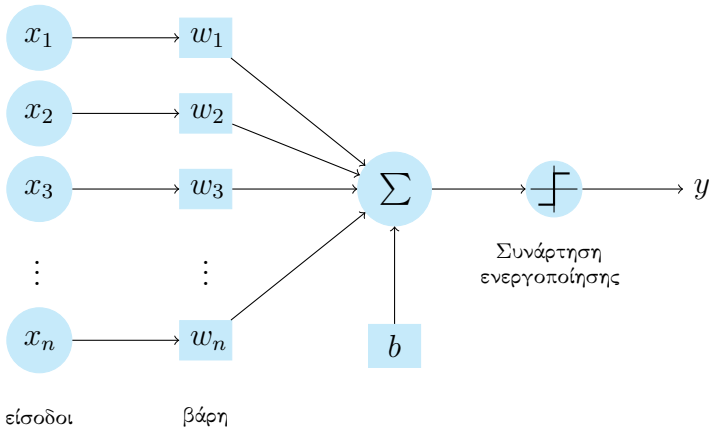
2.2.1 Πολυστρωματικά Perceptron

Τα πολυστρωματικά perceptron (multilayer perceptron - MLP) είναι μια κατηγορία τεχνητών νευρωνικών δικτύων (artificial neural networks - ANNs) πρόσθιας τροφοδότησης (feedforward).

Το πολυστρωματικό perceptron είναι ένα δίκτυο από νευρώνες που ονομάζονται perceptrons. Η βασική ιδέα του απλού perceptron προτάθηκε το 1958 από τον Rosenblatt [84]. Η είσοδος του perceptron είναι ένα διάνυσμα χαρακτηριστικών $\mathbf{x}$, ενώ η μοναδική έξοδος y υπολογίζεται μέσω του γινομένου του $\mathbf{x}$ με ένα διάνυσμα βαρών $\mathbf{w}$, στο οποίο προστίθεται η πόλωση (bias) b. Σε αυτό το αποτέλεσμα μπορεί προαιρετικά να χρησιμοποιηθεί μία μη γραμμική συνάρτηση g, η οποία ονομάζεται συνάρτηση ενεργοποίησης (activation function). Έτσι, η έξοδος y δίνεται από την παρακάτω εξίσωση:

$$y = g(\sum_{i=1}^{n} w_i x_i + b) = g(\mathbf{w}^T \mathbf{x} + b) \tag{2.1}$$

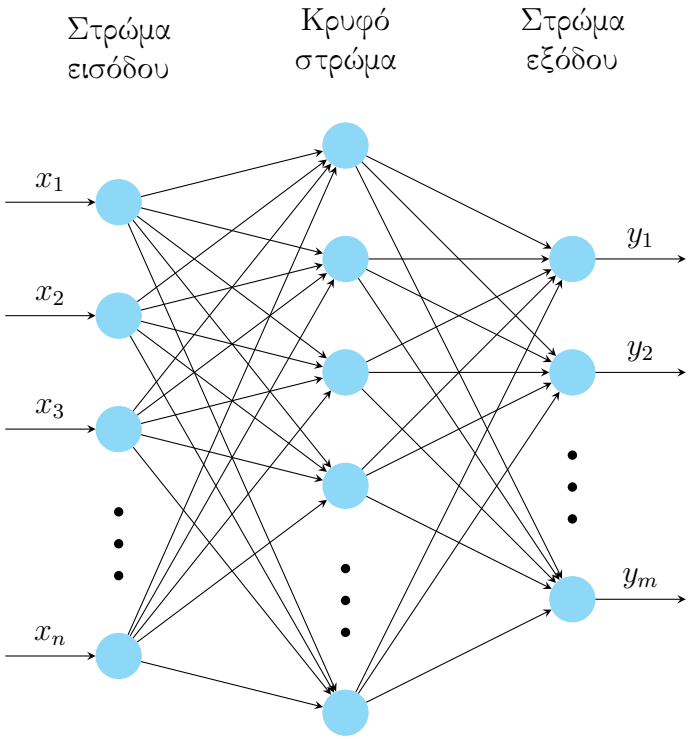
Το απλό perceptron απεικονίζεται γραφικά στο σχήμα 2.1. Αρχικά, είχε προταθεί η βηματική συνάρτηση ως συνάρτηση ενεργοποίησης. Επειδή, όμως, η βηματική συνάρτηση δεν είναι παραγωγίσιμη, πλέον επιλέγεται μία σιγμοειδής συνάρτηση. Οι δύο πιο συχνά χρησιμοποιούμενες συναρτήσεις είναι η λογιστική σιγμοειδής (logistic sigmoid) $1/(1+e^{-x})$ και η υπερβολική εφαπτομένη $tanh(x)$.



Σχήμα 2.1: Το απλό perceptron

Το πολυστρωματικό perceptron αποτελείται από τουλάχιστον τρία στρώματα, δηλαδή από ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Οι νευρώνες των διάφορων στρωμάτων είναι πλήρως συνδεδεμένοι, που σημαίνει ότι κάθε κόμβος ενός στρώματος συνδέεται με ένα συγκεκριμένο βάρος w_{ij} με κάθε κόμβο του επόμενου στρώματος. Εκτός

από τους κόμβους του στρώματος εισόδου, κάθε κόμβος χρησιμοποιεί μια μη γραμμική συνάρτηση ενεργοποίησης. Στο σχήμα 2.2 απεικονίζεται ένα πολυστρωματικό perceptron με ένα μόνο κρυφό στρώμα.



Σχήμα 2.2: Ένα πολυστρωματικό perceptron με ένα μόνο κρυφό στρώμα

Τα πολυστρωματικά perceptron χρησιμοποιούνται κυρίως σε προβλήματα επιβλεπόμενης μάθησης. Αυτό προϋποθέτει την ύπαρξη ενός συνόλου ζευγών εισόδων-εξόδων το οποίο χρησιμοποιείται για την εκπαίδευση, με βάση το οποίο πρέπει να προσαρμοστούν οι τιμές των παραμέτρων του δικτύου.

Για την εκπαίδευση του δικτύου, χρησιμοποιείται η τεχνική της οπισθοδιάδοσης (backpropagation) [89]. Ο αλγόριθμος αποτελείται από δύο βήματα. Στο ευθύ πέρασμα (forward pass), αποτιμάται η τιμή της εξόδου με βάση την αντίστοιχη είσοδο. Η έξοδος που προκύπτει διαφέρει από την επιθυμητή έξοδο, επομένως οι παράμετροι του δικτύου θα πρέπει να προσαρμοστούν έτσι ώστε να μειωθεί το σφάλμα.

Αυτό γίνεται με το δεύτερο βήμα, το ανάστροφο πέρασμα (reverse pass), όπου οι μερικές παράγωγοι μιας συνάρτησης σφάλματος ως προς τις παραμέτρους του δικτύου μεταφέρονται προς τα πίσω. Οι παράμετροι του δικτύου στη συνέχεια προσαρμόζονται χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης με βάση την κλίση. Η όλη διαδικασία επαναλαμβάνεται για κάθε ζεύγος

εισόδου-επιθυμητής εξόδου, και για έναν αριθμό εποχών, μέχρι οι τιμές των παραμέτρων να συγκλίνουν.

2.2.1.1 Ο αλγόριθμος οπισθοδιάδοσης

Έστω ένα σύνολο δεδομένων $X = \{(\vec{x}_1, \vec{y}_1), \ldots, (\vec{x}_N, \vec{y}_N)\}$ που αποτελείται από ζεύγη εισόδων $\vec{x}_i$ και επιθυμητών εξόδων $\vec{y}_i$ και ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης, του οποίου οι παράμετροι δηλώνονται συλλογικά ως θ. Οι παράμετροι αυτές είναι τα βάρη του δικτύου, όπου με w_{ij}^k συμβολίζεται το βάρος μεταξύ του κόμβου j του στρώματος k και του κόμβου i του στρώματος $k-1$, και οι πολώσεις των κόμβων, όπου με b_i^k ή w_{0i}^k συμβολίζεται η πόλωση του κόμβου i του στρώματος k. Δεν υπάρχουν συνδέσεις ανάμεσα σε κόμβους του ίδιου στρώματος και τα στρώματα είναι πλήρως συνδεδεμένα μεταξύ τους. Για τον ορισμό του σφάλματος μεταξύ της επιθυμητής εξόδου $\vec{y}_i$ και της υπολογισμένης από το δίκτυο εξόδου $\hat{\vec{y}}_i$, για μια δεδομένη είσοδο $\vec{x}_i$, ορίζεται μια συνάρτηση σφάλματος $E(X, \theta)$.

Η εκπαίδευση του δικτύου με τη χρήση της καθόδου κλίσης (gradient descent) προϋποθέτει τον υπολογισμό της μερικής παραγώγου της συνάρτησης σφάλματος $E(X, \theta)$ ως προς τα βάρη w_{ij}^k. Αν η είναι ο ρυθμός μάθησης και θ^t είναι οι παράμετροι του δικτύου στην επανάληψη t, τότε σε κάθε επανάληψη οι παράμετροι θα μεταβάλλονται ως εξής:

$$\theta^{t+1} = \theta^t - \eta \frac{\partial E(X, \theta)}{\partial \theta} \tag{2.2}$$

Στόχος είναι η εύρεση των βέλτιστων τιμών των παραμέτρων. Ένα μεγάλο πρόβλημα είναι ότι οι κόμβοι των κρυφών στρωμάτων δεν έχουν επιθυμητή τιμή εξόδου, επομένως δεν μπορεί να οριστεί συνάρτηση σφάλματος για αυτούς τους κόμβους. Αντίθετα, η συνάρτηση σφάλματος θα εξαρτάται από τις τιμές των παραμέτρων των προηγούμενων και των επόμενων στρωμάτων.

Ιστορικά, η συνάρτηση σφάλματος που χρησιμοποιήθηκε στην κλασική μέθοδο οπισθοδιάδοσης είναι η συνάρτηση μέσου τετραγωνικού σφάλματος:

$$E(X, \theta) = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{2.3}$$

όπου για λόγους απλότητας θεωρούμε ότι το δίκτυο έχει μόνο μία έξοδο (τα $\hat{y}_i$ και y_i δεν είναι πλέον διανύσματα). Προφανώς, μπορεί να επιλεγεί διαφορετική συνάρτηση σφάλματος.

Για τον υπολογισμό των μερικών παραγώγων γίνεται χρήση του κανόνα της αλυσίδας και του κανόνα του γινομένου του διαφορικού λογισμού. Η χρήση των

δύο αυτών κανόνων εξαρτάται από την παραγώγιση της συνάρτησης ενεργοποίησης. Για το λόγο αυτό, η βηματική συνάρτηση, η οποία δεν είναι συνεχής, άρα ούτε και παραγωγίσιμη, δεν χρησιμοποιείται πλέον ως συνάρτηση ενεργοποίησης.

Ο αλγόριθμος συνοψίζεται στα εξής βήματα:

1. Στο **ευθύ πέρασμα**, για κάθε ζεύγος εισόδου-εξόδου $(\vec{x}_d, y_d)$, υπολογίζονται και αποθηκεύονται οι τιμές των εξόδων $\hat{y}_d$, των ενεργοποιήσεων των κόμβων α_j^k και των εξόδων των κόμβων o_j^k, για κάθε κόμβο j κάθε στρώματος k, ξεκινώντας από το στρώμα 0 (το στρώμα εισόδου), και καταλήγοντας στο στρώμα m (το στρώμα εξόδου).

2. Στο **ανάστροφο πέρασμα**, για κάθε ζεύγος εισόδου-εξόδου $(\vec{x}_d, y_d)$, υπολογίζονται και αποθηκεύονται τα αποτελέσματα των $\frac{\partial E_d}{\partial w_{ij}^k}$, για κάθε βάρος w_{ij}^k που συνδέει τον κόμβο i του στρώματος $k-1$ με τον κόμβο j του στρώματος k, ξεκινώντας από το στρώμα m και καταλήγοντας στο στρώμα 0.

   (α′) Για το στρώμα εξόδου, υπολογίζεται ο όρος σφάλματος δ_1^m:

   $$\delta_1^m = (\hat{y_d} - y_d) g_o'(\alpha_1^m)$$

   όπου g_o είναι η συνάρτηση ενεργοποίησης του στρώματος εξόδου.

   (β′) Οι όροι σφάλματος των κρυφών στρωμάτων δ_j^k υπολογίζονται, ξεκινώντας από το στρώμα $m-1$ (τελευταίο κρυφό στρώμα) και προς τα πίσω:

   $$\delta_j^k = g'(\alpha_j^k) \sum_{l=1}^{r_{k+1}} w_{jl}^{k+1} \delta_l^{k+1}$$

   (γ′) Υπολογίζονται οι μερικές παράγωγοι των επιμέρους σφαλμάτων E_d ως προς τα βάρη w_{ij}^k:

   $$\frac{\partial E}{\partial w_{ij}^k} = \delta_j^k o_i^{k-1}$$

3. Οι μερικές παράγωγοι των σφαλμάτων για κάθε ζεύγος εισόδων-εξόδων συνδυάζονται ώστε να προκύψει η μερική παράγωγος της συνάρτησης σφάλματος για όλο το σύνολο δεδομένων:

   $$\frac{\partial E(X, \theta)}{\partial w_{ij}^k} = \frac{1}{N} \sum_{d=1}^{N} \frac{\partial E_d}{\partial w_{ij}^k}$$

4. Τα βάρη ενημερώνονται:

$$\Delta w_{ij}^k = -\eta \frac{\partial E(X, \theta)}{\partial w_{ij}^k}$$

2.2.2 Δένδρα Αποφάσεων

Τα δένδρα αποφάσεων (Decision Trees - DTs) είναι μια μη παραμετρική επιβλεπόμενη μέθοδος μάθησης. Στόχος είναι η δημιουργία ενός μοντέλου που να προβλέπει την τιμή μιας μεταβλητής μέσω της μάθησης απλών κανόνων απόφασης που προκύπτουν από τα χαρακτηριστικά των δεδομένων. Ένα δένδρο αποφάσεων χρησιμοποιείται για ταξινόμηση ή παλινδρόμηση, οπότε συνήθως αναφέρεται ως δένδρο ταξινόμησης ή δέντρο παλινδρόμησης, αντίστοιχα. Στην παρούσα εργασία, τα δένδρα αποφάσεων χρησιμοποιούνται αποκλειστικά για ταξινόμηση.

Ένα δένδρο αποφάσεων είναι ένας ταξινομητής που εκφράζεται ως αναδρομικός διαχωρισμός του χώρου παραδειγμάτων. Είναι μια κατευθυνόμενη δενδρική δομή, όπου η ρίζα δεν έχει καθόλου εισερχόμενες ακμές, ενώ όλοι οι υπόλοιποι κόμβοι έχουν ακριβώς μία εισερχόμενη ακμή. Ένας κόμβος που έχει εξερχόμενες ακμές αναφέρεται ως *εσωτερικός κόμβος* ή *κόμβος ελέγχου*. Οι υπόλοιποι κόμβοι ονομάζονται *φύλλα*, *τελικοί κόμβοι* ή *κόμβοι απόφασης*. Κάθε εσωτερικός κόμβος αντιπροσωπεύει έναν έλεγχο με βάση τις τιμές των χαρακτηριστικών εισόδου, ενώ κάθε ακμή αντιπροσωπεύει το αποτέλεσμα του ελέγχου. Στην πιο απλή περίπτωση, κάθε έλεγχος λαμβάνει υπ΄ όψιν την τιμή μόνο ενός χαρακτηριστικού. Επομένως, οι εσωτερικοί κόμβοι χωρίζουν τον χώρο παραδειγμάτων σε δύο ή περισσότερους υπο-χώρους. Τέλος, κάθε φύλλο αντιπροσωπεύει την τελική απόφαση ταξινόμησης, η οποία είναι είτε η επισήμανση μιας κλάσης, είτε ένα διάνυσμα πιθανοτήτων [83].

2.2.2.1 Κατασκευή δένδρων αποφάσεων

Η κατασκευή του βέλτιστου δυαδικού δένδρου αποφάσεων είναι NP-Complete πρόβλημα [49]. Μια απλή, άπληστη προσέγγιση για την κατασκευή δένδρων αποφάσεων περιγράφεται στη συνέχεια:

1. Επίλεξε το χαρακτηριστικό που πετυχαίνει τον καλύτερο διαχωρισμό μεταξύ των κατηγοριών, με βάση κάποιο κριτήριο.

2. Χώρισε τα δεδομένα σε υποσύνολα με βάση της τιμές του χαρακτηριστικού αυτού.

3. Για κάθε υποσύνολο που περιέχει περισσότερες από μία κατηγορίες, επανάλαβε τη διαδικασία.

4. Σταμάτησε εφόσον δεν υπάρχουν υποσύνολα που περιέχουν περισσότερες από μία κατηγορίες ή έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά.

Κάποιοι από τους πιο γνωστούς αλγόριθμους κατασκευής δένδρων αποφάσεων αναφέρονται στη συνέχεια:

- Ο **ID3** (Iterative Dichotomiser 3) [79] αναπτύχθηκε το 1986 από τον Ross Quinlan. Ο αλγόριθμος δημιουργεί ένα δέντρο πολλαπλών διαδρομών, επιλέγοντας με άπληστο τρόπο για κάθε κόμβο το κατηγορικό χαρακτηριστικό που θα αποφέρει το μεγαλύτερο κέρδος πληροφορίας (information gain) για τους κατηγορικούς στόχους. Τα δένδρα αναπτύσσονται στο μέγιστο μέγεθος τους και στη συνέχεια εφαρμόζεται ένα βήμα κλαδέματος (pruning) για τη βελτίωση της ικανότητας του δέντρου να γενικεύει σε άγνωστα δεδομένα.

- Ο αλγόριθμος **C4.5** [80], ο οποίος επίσης προτάθηκε από τον Quinlan, αποτελεί επέκταση του ID3. Ο C4.5 μπορεί να διαχειριστεί τόσο συνεχή, όσο και διακριτά χαρακτηριστικά. Για να το πετύχει αυτό, δημιουργεί ένα κατώφλι και διαχωρίζει τις εισόδους σε δυο σύνολα, ανάλογα με το αν η τιμή για το συγκεκριμένο χαρακτηριστικό είναι μεγαλύτερη από το κατώφλι ή όχι. Ο C4.5 μετατρέπει τα εκπαιδευμένα δένδρα σε σύνολα κανόνων if-then. Στη συνέχεια, αξιολογείται η ακρίβεια κάθε τέτοιου κανόνα για να προσδιοριστεί με ποια σειρά πρέπει να εφαρμοστούν. Τέλος, το κλάδεμα γίνεται με την αφαίρεση κλαδιών που δεν βελτιώνουν την απόδοση του μοντέλου και την αντικατάστασή τους από φύλλα.

- Ο **CART** (Classification and Regression Trees) [14] είναι παρόμοιος με τον C4.5, αλλά διαφέρει στο ότι υποστηρίζει αριθμητικές μεταβλητές (παλινδρόμηση) και δεν υπολογίζει σύνολα κανόνων. Ο CART κατασκευάζει δυαδικά δέντρα χρησιμοποιώντας το χαρακτηριστικό και το κατώφλι που αποφέρουν το μεγαλύτερο κέρδος πληροφορίας σε κάθε κόμβο.

2.2.2.2 Κριτήρια διαχωρισμού

Υπάρχουν διάφορα κριτήρια για το διαχωρισμό των εισόδων σε κάθε κόμβο, τα οποία συνήθως υπολογίζουν την ομοιογένεια της μεταβλητής-στόχου στα υποσύνολα διαχωρισμού. Αν p_{mk} είναι το ποσοστό της κλάσης k στον κόμβο m του δένδρου και X_m είναι τα δεδομένα εκπαίδευσης στον κόμβο m, τότε τα πιο γνωστά κριτήρια είναι τα εξής:

- Η μετρική **Gini Impurity** εκφράζει το πόσο συχνά ένα τυχαία επιλεγμένο στοιχείο από το σύνολο θα χαρακτηριζόταν λανθασμένα αν ήταν

τυχαία επισημασμένο σύμφωνα με την κατανομή επισημάνσεων στο υποσύνολο:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{2.4}$$

- Η **Εντροπία** είναι το μέτρο της ποσότητας της αβεβαιότητας ή τυχαιότητας στα δεδομένα:

$$H(X_m) = -\sum_k p_{mk} log(p_{mk}) \tag{2.5}$$

- Το **κέρδος πληροφορίας (information gain)** είναι μια έννοια που προέρχεται από τη θεωρία πληροφοριών, η οποία αναφέρεται στη μείωση του επιπέδου τυχαιότητας σε ένα σύνολο δεδομένων, επομένως βασίζεται στην έννοια της εντροπίας.

- Το **λάθος ταξινόμησης (misclassification error)** ορίζεται ως εξής:

$$H(X_m) = 1 - max(p_{mk}) \tag{2.6}$$

- Η **μείωση της διακύμανσης (variance reduction)** χρησιμοποιείται στις περιπτώσεις που η μεταβλητή-στόχος είναι συνεχής (δένδρα παλινδρόμησης).

2.2.3 Τυχαίο Δάσος

Τα τυχαία δάση (Random Forests) [44, 13] είναι μοντέλα μάθησης συνόλου ή συνδυαστικής μάθησης (ensemble learning), για την ταξινόμηση ή την παλινδρόμηση, που λειτουργούν κατασκευάζοντας ένα πλήθος δένδρων αποφάσεων, και δίνοντας ως έξοδο την πιο συχνή κλάση ή τη μέση πρόβλεψη των επιμέρους δένδρων. Τα τυχαία δάση διορθώνουν την τάση των δένδρων αποφάσεων να κάνουν υπερπροσαρμογή (overfitting) στο σύνολο εκπαίδευσης.

Τα τυχαία δάση προτάθηκαν αρχικά από τον Tin Kam Ho το 1995 [44]. Μια επέκταση του αλγορίθμου [13] υλοποιήθηκε από τους Leo Breiman και Adele Cutler, οι οποίοι και καθιέρωσαν τον όρο «Τυχαία Δάση» ως εμπορικό σήμα (trademark). Η επέκταση συνδυάζει την αρχική ιδέα για τυχαία επιλογή των χαρακτηριστικών, με την ιδέα του "bagging".

2.2.3.1 Bagging

Η πολλαπλή δειγματοθέτηση (bootstrap aggregating ή εναλλακτικά bagging), είναι ένας μετα-αλγόριθμος μάθησης συνόλου ο οποίος σχεδιάστηκε για

τη βελτίωση της σταθερότητας και της ακρίβειας των αλγορίθμων μηχανικής μάθησης. Επιπρόσθετα, μειώνει τη διακύμανση, ενώ βοηθά στην αποφυγή της υπερπροσαρμογής. Ο αλγόριθμος εκπαίδευσης των τυχαίων δασών εφαρμόζει τον αλγόριθμο bagging στα επιμέρους δένδρα. Παρ' όλα αυτά, το bagging μπορεί να εφαρμοστεί σε οποιαδήποτε μέθοδο μάθησης.

Δεδομένου ενός συνόλου εκπαίδευσης $X = x_1, \ldots, x_n$ με αποκρίσεις $Y = y_1, \ldots, y_n$, ο αλγόριθμος επαναληπτικά επιλέγει με επανατοποθέτηση ένα τυχαίο δείγμα από το σύνολο εκπαίδευσης και εκπαιδεύει τα δένδρα σε αυτά τα δείγματα:

Για $b = 1, \ldots, B$:

1. Επίλεξε, με επανατοποθέτηση, n δείγματα εκπαίδευσης X_b, Y_b από τα X, Y.

2. Εκπαίδευσε ένα δένδρο f_b στα δείγματα X_b, Y_b.

Μετά την εκπαίδευση, στην περίπτωση της παλινδρόμησης, οι προβλέψεις για τα μη γνωστά δείγματα x' προκύπτουν μέσω του υπολογισμού του μέσου όρου των προβλέψεων των επιμέρους δένδρων για το x', όπως δίνεται στην εξίσωση 2.7:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{2.7}$$

Στην περίπτωση της ταξινόμησης, το αποτέλεσμα προκύπτει από την επιλογή της κλάσης που αποτελεί πλειοψηφία στις επιμέρους προβλέψεις.

Η παραπάνω διαδικασία οδηγεί σε καλύτερη απόδοση, καθώς μειώνει τη διακύμανση του μοντέλου. Καθώς οι προβλέψεις ενός μόνο δέντρου είναι ευαίσθητες στο θόρυβο που υπάρχει στο σύνολο εκπαίδευσης, ο μέσος όρος πολλών δένδρων λύνει αυτό το πρόβλημα, αρκεί τα δένδρα να μην έχουν συσχέτιση. Η εκπαίδευση πολλαπλών δένδρων πάνω στα ίδια δεδομένα οδηγεί σε συσχέτιση (και συχνά και στην παραγωγή των ίδιων ακριβώς δένδρων), γεγονός που εξηγεί για ποιο λόγο γίνεται δειγματοληψία κατά την δημιουργία του κάθε συνόλου δεδομένων εκπαίδευσης.

2.2.3.2 Bagging σε τυχαία δάση

Η παραπάνω διαδικασία περιγράφει τον αλγόριθμο στην γενική περίπτωση. Στα τυχαία δάση, χρησιμοποιείται ένας τροποποιημένος αλγόριθμος μάθησης δένδρων, ο οποίος επιλέγει ένα υποσύνολο των χαρακτηριστικών σε κάθε υποψήφιο διαχωρισμό στη διαδικασία μάθησης. Ο λόγος που γίνεται αυτό είναι ότι αν ένα ή λίγα χαρακτηριστικά είναι πολύ ισχυροί παράγοντες πρόβλεψης, αυτά

τα χαρακτηριστικά θα επιλεγούν σε πολλά από τα δένδρα, με αποτέλεσμα τη δημιουργία συσχετισμένων ή και όμοιων δένδρων.

2.2.4 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) [21, 106], ή Δίκτυα Διανυσμάτων Υποστήριξης (Support Vector Networks), είναι μοντέλα επιβλεπόμενης μάθησης που αναλύουν δεδομένα με στόχο την ταξινόμηση ή την παλινδρόμηση. Τα SVM είναι δυαδικοί, μη-πιθανοτικοί (non-probabilistic) ταξινομητές. Δεδομένου ενός συνόλου παραδειγμάτων εκπαίδευσης, κάθε ένα εκ των οποίων ανήκει σε μία από δύο κατηγορίες, το SVM αναπαριστά τα δεδομένα σαν σημεία σε έναν πολυδιάστατο χώρο, στον οποίο κατασκευάζει ένα διαχωριστικό υπερεπίπεδο, το οποίο μεγιστοποιεί το κενό μεταξύ των δυο συνόλων παραδειγμάτων που αντιστοιχούν στις δύο κατηγορίες.

Τα SVM μπορούν να χρησιμοποιηθούν τόσο για γραμμικά, όσο και για μη γραμμικά διαχωρίσιμα δεδομένα. Κατά τη φάση της εκπαίδευσης, τα SVM αντιστοιχίζουν τις εισόδους σε ένα χώρο μεγάλης διάστασης, στην οποία ψάχνουν να βρουν υπερεπίπεδα τα οποία τις διαχωρίζουν. Το βέλτιστο υπερεπίπεδο, δηλαδή εκείνο το οποίο επιτυγχάνει το μέγιστο διαχωρισμό μεταξύ των κατηγοριών, ονομάζεται μέγιστο περιθώριο υπερεπίπεδο (maximum marginal hyperplane). Τα παραδείγματα με τη μικρότερη απόσταση από το μέγιστο περιθώριο υπερεπίπεδο καλούνται *διανύσματα υποστήριξης* (support vectors). Η χρήση των διανυσμάτων υποστήριξης καθιστά τα SVM αποδοτικά ως προς τη χρήση μνήμης, αφού τελικά χρησιμοποιούν μόνο ένα υποσύνολο των παραδειγμάτων εκπαίδευσης για την εύρεση του υπερεπιπέδου.

2.2.4.1 Γραμμικές Μηχανές Διανυσμάτων Υποστήριξης

2.2.4.1.1 Γραμμικά διαχωρίσιμα δεδομένα Έστω $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, l$, $y_i \in 1, -1$, $\mathbf{x} \in \mathbb{R}^d$ το σύνολο δεδομένων εκπαίδευσης, όπου τα διανύσματα $\mathbf{x}_i$ έχουν d διαστάσεις. Ζητούμενο είναι το υπερεπίπεδο το οποίο χωρίζει τα διανύσματα $\mathbf{x}_i$ για τα οποία $y_i = 1$, από τα διανύσματα $\mathbf{x}_i$ για τα οποία $y_i = -1$, όπου η απόσταση ανάμεσα στο υπερεπίπεδο και το κοντινότερο διάνυσμα $\mathbf{x}_i$ και για τις δύο κατηγορίες μεγιστοποιείται.

Κάθε υπερεπίπεδο μπορεί να δοθεί από την παρακάτω εξίσωση:

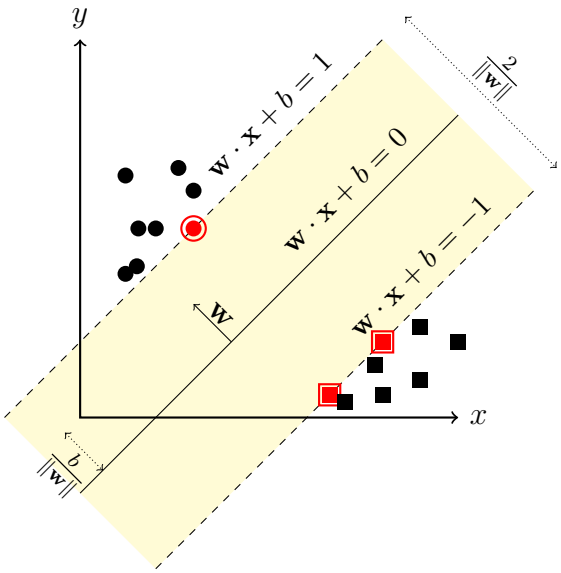
$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.8}$$

όπου $\mathbf{w}$ είναι το κάθετο (όχι απαραίτητα μοναδιαίο) διάνυσμα στο υπερεπίπεδο και b το κατώφλι.

Στην απλή περίπτωση, που τα δεδομένα είναι γραμμικά διαχωρίσιμα, μπορούμε να επιλέξουμε δύο παράλληλα υπερεπίπεδα, που διαχωρίζουν τις δύο κατηγορίες δεδομένων, έτσι ώστε η απόσταση ανάμεσά τους να είναι η μέγιστη δυνατή. Η περιοχή ανάμεσα στα δύο υπερεπίπεδα ονομάζεται περιθώριο (margin) και το υπερεπίπεδο μέγιστου περιθωρίου βρίσκεται ακριβώς στη μέση, ισαπέχει δηλαδή από τα δύο υπερεπίπεδα. Τα υπερεπίπεδα αυτά μπορούν να περιγραφούν από τις εξισώσεις:

$$\mathbf{w} \cdot \mathbf{x} + b = 1 \tag{2.9}$$

$$\mathbf{w} \cdot \mathbf{x} + b = -1 \tag{2.10}$$



Σχήμα 2.3: Το υπερεπίπεδο μέγιστου περιθωρίου, το περιθώριο και τα διανύσματα υποστήριξης ενός SVM που έχει εκπαιδευτεί να διαχωρίζει τα παραδείγματα δύο κλάσεων.

Η απόσταση ανάμεσα στα υπερεπίπεδα είναι ίση με $\frac{2}{\|\mathbf{w}\|}$, άρα για να μεγιστοποιηθεί η απόσταση πρέπει να ελαχιστοποιηθεί το $\|\mathbf{w}\|$.

Ένας επιπλέον περιορισμός είναι ότι τα διανύσματα που αντιστοιχούν στα δεδομένα δεν πρέπει να βρίσκονται εντός του περιθωρίου, άρα για κάθε i πρέπει:

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq 1, \text{ αν } y_i = 1 \tag{2.11}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1, \text{ αν } y_i = -1 \tag{2.12}$$

Οι εξισώσεις 2.11 και 2.12 μπορούν να ξαναγραφούν ως:

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \, \forall i \tag{2.13}$$

Η λύση για την περίπτωση των δύο διαστάσεων αναμένεται να έχει τη μορφή που φαίνεται στην εικόνα 2.3, όπου με κόκκινο απεικονίζονται τα διανύσματα υποστήριξης.

Το παραπάνω πρόβλημα μπορεί να εκφραστεί με τη χρήση μιας λαγκρανζιανής διατύπωσης (Lagrangian formulation) [16]. Αυτό γίνεται για δύο λόγους, ότι οι περιορισμοί της εξίσωσης 2.13 αντικαθίστανται από περιορισμούς των πολλαπλασιαστών Lagrange και ότι με την αναδιατύπωση του προβλήματος, τα δεδομένα εκπαίδευσης θα εμφανίζονται μόνο με τη μορφή εσωτερικών γινομένων μεταξύ διανυσμάτων.

Έστω ότι α_i, $i = 1, ..., l$ είναι οι θετικοί πολλαπλασιαστές Lagrange, ένας για κάθε περιορισμό της εξίσωσης 2.13. Η Λαγκρανζιανή που προκύπτει είναι η παρακάτω:

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x_i} \cdot \mathbf{w} - b) + \sum_{i=1}^{l} \alpha_i \tag{2.14}$$

Στόχος είναι η ελαχιστοποίηση της L_p ως προς τα $\mathbf{w}$, b, με την απαίτηση ταυτόχρονα ότι οι παράγωγοι της L_p ως προς τα α_i θα γίνουν ίσες με μηδέν και $\alpha_i \geq 0$. Αυτό είναι ένα πρόβλημα κυρτού τετραγωνικού προγραμματισμού, που σημαίνει ότι μπορεί να λυθεί το «δυαδικό» πρόβλημα: η μεγιστοποίηση της L_p, υπό τον περιορισμό ότι η κλίση της L_p ως προς τα $\mathbf{w}$, b μηδενίζεται. Αυτό συμβαίνει όταν:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{2.15}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.16}$$

Με την αντικατάσταση αυτών των περιορισμών στην εξίσωση 2.14, δεν υπάρχει πλέον εξάρτηση από τα $\mathbf{w}$ και b:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.17}$$

Επομένως, η εκπαίδευση των διανυσμάτων υποστήριξης, για την περίπτωση των γραμμικά διαχωρίσιμων δεδομένων, αποτελείται από τη μεγιστοποίηση της L_D ως προς τα α_i, υπο τους περιορισμούς της εξίσωσης 2.16 και $\alpha_i \geq 0$. Υπάρχει ένας πολλαπλασιαστής Lagrange α_i για κάθε σημείο δεδομένων, και για όσα σημεία αποτελούν διανύσματα υποστήριξης ισχύει ότι $\alpha_i > 0$ (για τα υπόλοιπα σημεία ισχύει $\alpha_i = 0$).

2.2.4.1.2 Μη γραμμικά διαχωρίσιμα δεδομένα

Όταν τα δεδομένα δεν είναι διαχωρίσιμα, θα πρέπει να χαλαρώσουν οι περιορισμοί των εξισώσεων 2.11 και 2.12, με την εισαγωγή των θετικών μεταβλητών ξ_i, $i = 1, \ldots, l$, οι οποίες ονομάζονται μεταβλητές χαλαρότητας (slack variables). Οι περιορισμοί πλέον γίνονται:

$$\mathbf{x_i} \cdot \mathbf{w} + b \geq 1 - \xi_i, \text{ αν } y_i = 1 \tag{2.18}$$

$$\mathbf{x_i} \cdot \mathbf{w} + b \leq -1 + \xi_i, \text{ αν } y_i = -1 \tag{2.19}$$

$$\xi_i \geq 0 \ \forall i \tag{2.20}$$

Για να συμβεί μια λανθασμένη ταξινόμηση, το αντίστοιχο ξ_i ξεπερνάει τη μονάδα, έτσι το $\sum_i \xi_i$ είναι το ανώτερο όριο στον αριθμό των σφαλμάτων ταξινόμησης. Η αντικειμενική συνάρτηση που πρέπει να ελαχιστοποιηθεί αλλάζει από την $||\mathbf{w}||^2/2$ στην $||\mathbf{w}||^2/2 + C(\sum_i \xi_i)^k$, όπου C είναι μία παράμετρος που ορίζει το βάρος του κόστους των λανθασμένων ταξινομήσεων, ενώ συνήθως επιλέγεται η τιμή $k = 1$ ώστε το πρόβλημα να είναι τετραγωνικό.

Η Λαγκρανζιανή συνάρτηση του αρχικού προβλήματος είναι η εξής:

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\{y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i \tag{2.21}$$

όπου οι μ_i είναι πολλαπλασιαστές Lagrange, οι οποίοι εισάγονται για να ενισχύσουν το γεγονός ότι τα ξ_i είναι θετικά.

2.2.4.2 Μη γραμμικές Μηχανές Διανυσμάτων Υποστήριξης

Για τη μη γραμμική περίπτωση, οι Boser et al. [11] έδειξαν ότι ένα «κόλπο» που είχε προταθεί παλαιότερα [2] μπορεί να χρησιμοποιηθεί για τη γενίκευση των μεθόδων που περιγράφηκαν προηγουμένως. Έστω ότι τα δεδομένα εκπαίδευσης μετασχηματίζονται σε έναν άλλο Ευκλείδιο χώρο $\mathcal{H}$ (πιθανώς απείρων διαστάσεων), χρησιμοποιώντας μια απεικόνιση Φ, τέτοια ώστε $\Phi : \mathbb{R}^d \mapsto \mathcal{H}$. Υποθέτουμε ότι τα μετασχηματισμένα πλέον σημεία $\Phi(\mathbf{x}_i)$ μπορούν να διαχωριστούν γραμμικά.

Αρχικά, τα δεδομένα στο πρόβλημα εκπαίδευσης εμφανίζονται με τη μορφή εσωτερικών γινομένων, δηλαδή $\mathbf{x}_i \cdot \mathbf{x}_j$. Μετά το μετασχηματισμό, ο αλγόριθμος θα βασίζεται μόνο σε εσωτερικά γινόμενα του $\mathcal{H}$, δηλαδή συναρτήσεις της μορφής $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Αν μπορούσε να βρεθεί μία συνάρτηση K, τέτοια ώστε $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, τότε η εκπαίδευση του μοντέλου μπορεί να γίνει μόνο με τη χρήση της K, χωρίς να γνωρίζουμε επακριβώς τι είναι η Φ. Οι συναρτήσεις αυτής της μορφής ονομάζονται συναρτήσεις πυρήνα (kernel functions).

Σύμφωνα με τη συνθήκη Mercer [105, 22], υπάρχουν ένας χώρος $\mathcal{H}$ και μια απεικόνιση Φ για κάποια συνάρτηση πυρήνα K, αν και μόνο αν, για κάθε $g(\mathbf{x})$, τέτοια ώστε το $\int g(\mathbf{x})^2 d\mathbf{x}$ να είναι πεπερασμένο, ισχύει ότι:

$$\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0 \tag{2.22}$$

Γενικά, είναι δύσκολο να αποδειχτεί ότι η συνθήκη Mercer ικανοποιείται για μια συνάρτηση πυρήνα.

Οι κυριότερες συναρτήσεις πυρήνα που χρησιμοποιούνται συχνά είναι:

- Η πολυωνυμική:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p \text{ ή } K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^p \tag{2.23}$$

- Η Γκαουσσιανή RBF (radial basis function):

$$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}} \tag{2.24}$$

- Η σιγμοειδής:

$$K(\mathbf{x}, \mathbf{y}) = tanh(\kappa \mathbf{x} \cdot \mathbf{y} - \delta) \tag{2.25}$$

2.3 Μοντέλα μη επιβλεπόμενης μάθησης

2.3.0.1 Αλγόριθμος Διάδοσης Συνάφειας

Η Διάδοση Συνάφειας (Affinity Propagation) [32] είναι ένας αλγόριθμος συσταδοποίησης που βασίζεται στην ανταλλαγή μηνυμάτων μεταξύ σημείων δεδομένων. Ο αλγόριθμος αναγνωρίζει υποδείγματα/πρότυπα (exemplars) ανάμεσα στα σημεία δεδομένων και σχηματίζει συστάδες (clusters) γύρω από αυτά τα πρότυπα. Θεωρεί ταυτόχρονα όλα τα σημεία ως πιθανά πρότυπα και επαναληπτικά ανταλλάσσει μηνύματα μεταξύ των σημείων μέχρι να προκύψει ένα καλό σύνολο προτύπων και ομάδων. Ο αριθμός των συστάδων που προκύπτουν δεν είναι προκαθορισμένος, αλλά προκύπτει από τα δεδομένα που δίνονται ως είσοδος.

Έστω $x_1, \ldots, x_n$ ένα σύνολο σημείων δεδομένων, και s μία συνάρτηση που ποσοτικοποιεί την ομοιότητα μεταξύ δύο σημείων, τέτοια ώστε αν το x_i είναι πιο όμοιο με το x_j απ΄ ότι με το x_k, τότε $s(x_i, x_j) > s(x_i, x_k)$.

Τα μηνύματα μεταξύ των σημείων είναι δύο διαφορετικών κατηγοριών, και για κάθε μία ορίζεται ένας πίνακας, οι τιμές του οποίου ενημερώνονται σε κάθε βήμα του αλγορίθμου. Επομένως, οι πίνακες που ορίζονται είναι:

- Ο πίνακας **R** ("responsibility"), όπου κάθε τιμή $r(i,k)$ υποδεικνύει πόσο κατάλληλο είναι το σημείο x_k για να αποτελέσει υπόδειγμα για το σημείο x_i, σε σχέση με τα άλλα υποψήφια υποδείγματα για το x_i.

- Ο πίνακας **A** ("availability"), όπου κάθε τιμή $\alpha(i,k)$ αντιπροσωπεύει το πόσο «κατάλληλο» θα ήταν για το x_i να επιλέξει το x_k ως υπόδειγμα, λαμβάνοντας υπ΄ όψιν την προτίμηση των άλλων σημείων για το x_k ως υπόδειγμα.

Οι πίνακες **R** και **A** αρχικοποιούνται με όλα τα στοιχεία τους ίσα με το μηδέν. Στη συνέχεια, ο αλγόριθμος εκτελεί τις ακόλουθες ενημερώσεις επαναληπτικά:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{\alpha(i,k') + s(i,k')\} \tag{2.26}$$

$$\alpha(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\right), \text{ αν } i \neq k \tag{2.27}$$

$$\alpha(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)) \tag{2.28}$$

Ο αλγόριθμος τερματίζει εάν οι αποφάσεις για τα πρότυπα και τα όρια των συστάδων παραμένουν αμετάβλητες για μια σειρά επαναλήψεων ή εάν επιτευχθεί ο μέγιστος αριθμός επαναλήψεων. Τα υποδείγματα εξάγονται από τους τελικούς πίνακες, ως τα στοιχεία για τα οποία ισχύει $r(i,i) + \alpha(i,i) > 0$.

2.4 Μετρικές αξιολόγησης

Σε αυτή την ενότητα θα παρουσιάσουμε τις πιο γνωστές μετρικές για την αξιολόγηση των ταξινομητών. Αρχικά, θα πρέπει να οριστούν οι όροι *αληθή θετικά* (true positives - tp), *αληθή αρνητικά* (true negatives - tn), *ψευδή θετικά* (false positives - fp), και *ψευδή αρνητικά* δείγματα (false negatives - fn). Στην δυαδική ταξινόμηση έχουμε δύο κλάσεις, οι οποίες συνήθως ονομάζονται θετική και αρνητική κλάση. Επομένως, οι όροι *θετικά* και *αρνητικά* αναφέρονται στην πρόβλεψη του ταξινομητή σε σχέση με την κλάση στην οποία ανήκει το κάθε δείγμα, ενώ οι όροι *αληθή* και *ψευδή* αναφέρονται στο κατά πόσο η συγκεκριμένη πρόβλεψη αντιστοιχεί στην εξωτερική κρίση (στην παρατήρηση).

Η *ορθότητα* (accuracy) είναι ο λόγος των σωστά ταξινομημένων δειγμάτων προς τον συνολικό αριθμό δειγμάτων:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.29}$$

Η ορθότητα έχει το μειονέκτημα ότι δεν έχει καλές επιδόσεις σε μη ισορροπημένα σύνολα δεδομένων. Αν, για παράδειγμα, έχουμε ένα σύνολο δεδομένων με 95 αρνητικά και 5 θετικά δείγματα, ένα σύστημα το οποίο ταξινομεί όλα τα δείγματα ως αρνητικά θα έχει ορθότητα ίση με 0.95. Για το λόγο αυτό, συνήθως επιλέγεται να χρησιμοποιούνται η ακρίβεια, η ανάκληση και το F1 Score.

Η *ακρίβεια* (precision) είναι ο λόγος των αληθών θετικών δειγμάτων προς τον αριθμό των αληθών θετικών συν τον αριθμό των ψευδών θετικών:

$$Precision = \frac{tp}{tp + fp} \tag{2.30}$$

Η ακρίβεια είναι ο λόγος των σωστών αποτελεσμάτων προς τον συνολικό αριθμό αποτελεσμάτων που επιστράφηκαν. Διαισθητικά, δείχνει την ικανότητα του ταξινομητή να επιστρέφει μόνο συναφή δείγματα.

Η *ανάκληση* (recall) είναι ο λόγος των αληθών θετικών δειγμάτων προς τον αριθμό των αληθών θετικών συν τον αριθμό των ψευδών αρνητικών:

$$Recall = \frac{tp}{tp + fn} \tag{2.31}$$

Επομένως, η ανάκληση είναι το ποσοστό των σωστών αποτελεσμάτων προς τον συνολικό αριθμό αποτελεσμάτων που θα έπρεπε να έχουν επιστραφεί, άρα δείχνει την ικανότητα του ταξινομητή να βρίσκει όλα τα συναφή δείγματα.

Ιδανικά, θα θέλαμε να έχουμε υψηλές τιμές τόσο για την ακρίβεια, όσο και για την ανάκληση. Συνήθως, όμως, μεταξύ της ακρίβειας και της ανάκλησης υπάρχει ένας συμβιβασμός (trade-off). Ένα μέτρο το οποίο συνδυάζει την ακρίβεια και την ανάκληση είναι το F1 Score (ή F-score ή F-measure), το οποίο υπολογίζεται από τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.32}$$

Στην περίπτωση της ταξινόμησης πολλαπλών κλάσεων (όπου υπάρχουν περισσότερες από δύο κλάσεις, αλλά κάθε δείγμα μπορεί να ανήκει σε μία μόνο κλάση), θα πρέπει τα F1 Scores των επιμέρους κλάσεων να συνδυαστούν με κάποιον τρόπο.

Ο πρώτος τρόπος είναι να υπολογιστεί ο μέσος όρος των F1 Scores των επιμέρους κλάσεων. Η μετρική αυτή ονομάζεται macro-averaged F1-score, ή macro-F1.

Κατά τον υπολογισμό του macro-F1, αναθέτουμε ίσα βάρη σε όλες τις κλάσεις. Με τη χρήση του weighted-average F1-score, ή weighted-F1, κατά τον υπολογισμό του μέσου όρου, κάθε κλάση έχει βάρος ανάλογο με τον αριθμό των δειγμάτων που περιέχει.

Η τελευταία παραλλαγή είναι το micro-averaged F1-score, ή micro-F1. Σε αυτή την περίπτωση, αθροίζουμε τα αληθή θετικά, τα ψευδή θετικά και τα ψευδή αρνητικά δείγματα για όλες τις κλάσεις, και με βάση αυτά υπολογίζουμε πρώτα την ακρίβεια και την ανάκληση (οι οποίες πλέον είναι οι micro-precision και micro-recall, αντίστοιχα). Από την ακρίβεια και την ανάκληση προκύπτει μετά το micro-F1.

Κεφάλαιο 3

Επεξεργασία Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας (Natural language processing ή NLP) είναι ένα διεπιστημονικό πεδίο, το οποίο ασχολείται με την αλληλεπίδραση μεταξύ ηλεκτρονικών υπολογιστών και ανθρώπινων/φυσικών γλωσσών. Αποτελεί υπο-πεδίο της γλωσσολογίας, της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης. Η πλευρά της γλωσσολογίας ασχολείται με την ανάλυση της γλώσσας, του σχηματισμού της, της σύνταξης, της σημασίας και του πλαισίου μέσα στο οποίο βρίσκεται. Η πλευρά της επιστήμης υπολογιστών ασχολείται με την εφαρμογή αυτών των γλωσσικών γνώσεων για τη δημιουργία αλγορίθμων, με τη βοήθεια υπο-πεδίων όπως η τεχνητή νοημοσύνη και η μηχανική μάθηση.

Ως γνωστόν, οι υπολογιστές καταλαβαίνουν αριθμούς, και όχι χαρακτήρες, λέξεις ή προτάσεις. Επομένως, ένα σημαντικό κομμάτι κάθε συστήματος επεξεργασίας φυσικής γλώσσας είναι η μετατροπή των μη δομημένων δεδομένων (τα οποία μπορεί να είναι στη μορφή γραπτού κειμένου ή ομιλίας), σε μια αναπαράσταση την οποία μπορεί να καταλαβαίνει ο υπολογιστής.

Ιστορικά, τα πρώτα συστήματα επεξεργασίας φυσικής γλώσσας ήταν συστήματα κανόνων, δηλαδή συστήματα στα οποία κωδικοποιούνταν ένα σύνολο κανόνων με το χέρι. Αργότερα, εξαιτίας της ανάγκης για αυτόματη εξαγωγή αυτών των κανόνων, έγινε χρήση στατιστικών μεθόδων και μηχανικής μάθησης, πράγμα που συνεχίζεται μέχρι και σήμερα. Επομένως, τα συστήματα επεξεργασίας φυσικής γλώσσας βασίζονται σε μεγάλο βαθμό στους αλγόριθμους μηχανικής μάθησης και στα στατιστικά μοντέλα.

Σε αυτό το κεφάλαιο θα αναλυθούν διάφορα θέματα της Επεξεργασίας Φυσικής Γλώσσας και της Εξόρυξης Κειμένου, τα οποία αποτελούν το θεωρητικό υπόβαθρο για τα επόμενα κεφάλαια. Πιο συγκεκριμένα, θα αναφερθούμε στην προεπεξεργασία του κειμένου, ενώ θα αναλυθούν διάφορες μέθοδοι αναπαράστασης και κωδικοποίησης των εγγράφων και των λέξεων, όπως είναι το σχήμα

tf-idf, η θεματική μοντελοποίηση και οι διανυσματικές παραστάσεις λέξεων.

3.1 Προεπεξεργασία Κειμένου

Μια πολύ σημαντική διαδικασία στην εξόρυξη γνώσης από κείμενο (text mining), την επεξεργασία φυσικής γλώσσας και την ανάκτηση πληροφορίας (information retrieval) είναι η προεπεξεργασία του κειμένου (text preprocessing). Ο στόχος της προεπεξεργασίας είναι η εξαγωγή χρήσιμης και δομημένης πληροφορίας από μη δομημένα έγγραφα κειμένου.

Με απλά λόγια, η προεπεξεργασία κειμένου είναι η μετατροπή του κειμένου σε μία μορφή που είναι αναλύσιμη για το εκάστοτε πρόβλημα. Η προεπεξεργασία κειμένου μπορεί να οδηγήσει σε μείωση του όγκου των δεδομένων, γεγονός που συνήθως αυξάνει την αποτελεσματικότητα. Η ιδανική προεπεξεργασία κειμένου είναι διαφορετική για κάθε σύστημα, και εξαρτάται από το πρόβλημα που πρέπει να λυθεί, καθώς και από την προσέγγιση που ακολουθείται.

Τα πιο γνωστά βήματα της προεπεξεργασίας δίνονται στη συνέχεια:

- Μετατροπή κεφαλαίων γραμμάτων σε πεζά: Είναι μία από τις πιο απλές και αποτελεσματικές μορφές προεπεξεργασίας. Είναι εφαρμόσιμη σε όλα σχεδόν τα προβλήματα εξόρυξης κειμένου και επεξεργασίας φυσικής γλώσσας. Συμβάλλει σημαντικά στη συνοχή της αναμενόμενης εξόδου, αφού διάφορες παραλλαγές της ίδιας λέξης (που είναι γραμμένες με το πρώτο γραμμα κεφαλαίο, με όλα τα γράμματα μικρά ή με όλα τα γράμματα κεφαλαία) αντιστοιχίζονται στην ίδια λέξη μετά τη μετατροπή.

- Ανίχνευση προτάσεων: Χωρίζει το κείμενο σε προτάσεις, με τη βοήθεια των σημείων στίξης.

- Ανίχνευση λεκτικών μονάδων (tokenization): Η διαδικασία παίρνει σαν είσοδο ένα κείμενο και επιστρέφει τις μεμονωμένες λέξεις.

- Ορθογραφικός έλεγχος: Μετατρέπει λέξεις που έχουν γραφτεί με λάθος τρόπο στην κανονική τους μορφή. Είναι ιδιαίτερα χρήσιμος σε κείμενα που περιέχουν θόρυβο, όπως π.χ. σχόλια στα κοινωνικά δίκτυα ή γραπτά μηνύματα.

- Αποκοπή καταλήξεων ή αναγωγή στο θέμα (stemming): Η διαδικασία επιστέφει κάθε λέξη χωρίς την κατάληξή της. Αυτό έχει σαν αποτέλεσμα ότι λέξεις με την ίδια ρίζα τελικά αντιστοιχίζονται στην ίδια λέξη.

- Λημματοποίηση (lemmatization): Η λημματοποίηση επιστρέφει για κάθε λέξη το αντίστοιχο λήμμα. Ως διαδικασία έχει τον ίδιο σκοπό που έχει και η αποκοπή καταλήξεων, παρ΄ όλα αυτά είναι πιο πολύπλοκη, αφού δεν αφαιρεί απλά την κατάληξη της λέξης, αλλά βρίσκει τη ρίζα από την οποία προέρχεται.

- Αφαίρεση αλφαριθμητικών (alphanumerics)

- Αφαίρεση ειδικών χαρακτήρων και σημείων στίξης

- Αφαίρεση των συχνών λέξεων (stopwords): Οι συχνές λέξεις δεν συμβάλλουν στο νόημα μίας πρότασης, επομένως μπορούν να αφαιρεθούν χωρίς να αλλάζει το νόημα της πρότασης. Ένα επιπλέον πλεονέκτημα της αφαίρεσης των συχνών λέξεων είναι ότι μειώνεται το λεξιλόγιο, γεγονός που αυξάνει την απόδοση των μετέπειτα βημάτων.

- Επισήμανση των λέξεων ως μερών του λόγου (part-of-speech tagging): Προσθέτει σε κάθε λέξη την αντίστοιχη επισήμανση, ανάλογα με την κατηγορία από μέρη του λόγου στην οποία ανήκει.

- Κατασκευή ευρετηρίου

Όπως εξηγήθηκε και νωρίτερα, ένα ή περισσότερα από αυτά τα βήματα μπορούν να παραληφθούν. Για παράδειγμα, η αποκοπή καταλήξεων και η λημματοποίηση είναι παρόμοιες διαδικασίες, επομένως επιλέγεται να γίνει μόνο η μία από τις δύο.

3.2 Το σχήμα tf-idf

Στην ανάκτηση πληροφορίας, το σχήμα tf-idf [90] (συντομογραφία του term frequency-inverse document frequency) είναι μια αριθμητική στατιστική που απεικονίζει τη σημασία μιας λέξης για ένα έγγραφο σε μια συλλογή ή ένα σώμα κειμένων. Η τιμή tf-idf αυξάνεται αναλογικά με το πόσες φορές εμφανίζεται μια λέξη στο έγγραφο, ενώ αντισταθμίζεται από τον αριθμό των εγγράφων στη συλλογή που περιέχουν τη λέξη. Προκύπτει ως το γινόμενο δυο όρων, του όρου tf (term frequency) και του όρου idf (inverse document frequency):

- Ο όρος **tf** μετρά πόσο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο. Δεδομένου ότι κάθε έγγραφο έχει διαφορετικό μήκος, είναι πιθανό ότι ένας όρος θα εμφανίζεται πολύ περισσότερες φορές σε μεγάλα έγγραφα παρά σε μικρότερα. Επομένως, κανονικοποιούμε διαιρώντας με το μήκος του εγγράφου (που είναι ο συνολικός αριθμός όρων στο έγγραφο). Αν

$f_{t,d}$ είναι ο αριθμός των εμφανίσεων του όρου t στο έγγραφο d, τότε το tf δίνεται από την εξίσωση:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{3.1}$$

- Αντίστοιχα, το **idf** μετρά πόσο σημαντικός είναι ένας όρος. Κατά τον υπολογισμό του tf, όλοι οι όροι θεωρούνται εξίσου σημαντικοί. Ωστόσο, ορισμένες λέξεις μπορεί να εμφανίζονται πολλές φορές, χωρίς να έχουν μεγάλη σημασία (π.χ. τα stopwords). Αντίθετα, άλλες λέξεις μπορεί να εμφανίζονται μόνο σε λίγα έγγραφα επειδή αποτελούν ειδικό λεξιλόγιο. Επομένως, πρέπει να μειώσουμε το βάρος στους συχνούς όρους, και να το αυξήσουμε στις σπάνιες λέξεις. Το idf δίνεται από την εξίσωση που ακολουθεί, όπου $\mathcal{D}$ είναι η συλλογή εγγράφων:

$$idf(t,\mathcal{D}) = log\left(\frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : t \in d\}|}\right) \tag{3.2}$$

Τα βάρη tf-idf τελικά προκύπτουν ως εξής:

$$tfidf(t,d,\mathcal{D}) = tf(t,d) \cdot idf(t,\mathcal{D}) \tag{3.3}$$

Μια μεγάλη τιμή του βάρους tf-idf αντιπροσωπεύει έναν όρο με υψηλή συχνότητα εμφάνισης στο συγκεκριμένο έγγραφο, αλλά χαμηλή συχνότητα εμφάνισης του όρου σε ολόκληρη τη συλλογή. Όπως εξηγήθηκε και νωρίτερα, τα βάρη τείνουν να φιλτράρουν τους κοινούς όρους.

Τα βάρη αυτά αποτελούν μια αριθμητική αναπαράσταση των εγγράφων, η οποία μπορεί να χρησιμοποιηθεί για τη σύγκριση των εγγράφων μεταξύ τους, και βρίσκει εφαρμογή στην αναζήτηση, στην ταξινόμηση και την ομαδοποίηση εγγράφων.

Το tf-idf ακολουθεί τη φιλοσοφία του μοντέλου bag-of-words, όπου ένα κείμενο αντιπροσωπεύεται από το σύνολο των λέξεων που περιέχει, αγνοώντας τη γραμματική και τη σειρά των λέξεων, διατηρώντας όμως την πολλαπλότητα.

3.3 Μοντελοποίηση Θεμάτων

Η μοντελοποίηση θεμάτων ή θεματική μοντελοποίηση (topic modelling) είναι μια κατηγορία αλγορίθμων που έχουν ως σκοπό την ανακάλυψη της θεματικής πληροφορίας σε μεγάλα αρχεία εγγράφων. Αποτελούν στατιστικές μεθόδους, οι οποίες αναλύουν τις λέξεις των αρχικών κειμένων για να ανακαλύψουν τη θεματολογία που υπάρχει σε αυτά, πως η θεματολογία συνδέει τα

κείμενα μεταξύ τους, και πως αλλάζει κατά τη διάρκεια του χρόνου [8]. Οι αλγόριθμοι θεματικής μοντελοποίησης δεν απαιτούν την ύπαρξη επισημάνσεων (ή ετικετών - labels) στα έγγραφα, αλλά τα θέματα προκύπτουν από την ανάλυση των αρχικών κειμένων.

Στη συνέχεια θα αναφερθούμε στο πιο γνωστό μοντέλο θέματος (topic model), που είναι η Λανθάνουσα Ανάθεση Dirichlet (Latent Dirichlet allocation - LDA) [9]. Το μοντέλο βασίζεται στη Λανθάνουσα Σημασιολογική Δεικτοδότηση (Latent Semantic Indexing - LSI) [24] και στην πιθανοτική Λανθάνουσα Σημασιολογική Δεικτοδότηση (probabilistic LSI) [46], ενώ έχει αποτελέσει εφαλτήριο για πολλά άλλα μοντέλα θέματος.

3.3.1 Λανθάνουσα Ανάθεση Dirichlet (Latent Dirichlet allocation - LDA)

Η λανθάνουσα ανάθεση Dirichlet (Latent Dirichlet allocation - LDA) είναι ένα γενετικό πιθανοτικό μοντέλο ενός σώματος εγγράφων, το οποίο βασίζεται στην ιδέα ότι κάθε έγγραφο είναι ένα μίγμα από θεματικές ενότητες (ή θέματα), όπου κάθε θεματική ενότητα αποτελείται από λέξεις με μια τυχαία κατανομή πιθανότητας [9]. Το μοντέλο LDA προτάθηκε από τους David Blei, Andrew Ng και Michael I. Jordan το 2003, και αποτελεί παράδειγμα θεματικής μοντελοποίησης.

Η βασική ιδέα πίσω από το μοντέλο είναι ότι κάθε έγγραφο προκύπτει από πολλαπλά θέματα, όπου ένα θέμα ορίζεται ως μια κατανομή πάνω σε ένα ορισμένο λεξιλόγιο όρων. Με απλά λόγια, μία συλλογή εγγράφων σχετίζεται με K θέματα, τα οποία εμφανίζονται σε κάθε έγγραφο σε διαφορετικές αναλογίες. Αυτή είναι μια παραδοχή που γίνεται συχνά, καθώς τα έγγραφα μιας συλλογής τείνουν να είναι ετερογενή, συνδυάζοντας ένα υποσύνολο βασικών ιδεών ή θεμάτων.

Η πρόκληση είναι ότι αυτά τα θέματα δεν είναι γνωστά εκ των προτέρων. Στόχος είναι η εύρεση αυτών των θεμάτων από τα δεδομένα με μη επιβλεπόμενο τρόπο. Με τη χρήση του LDA σε μια συλλογή κειμένων λαμβάνουμε ένα σύνολο θεμάτων, κατ΄ επέκταση την κατανομή των λέξεων ή κλειδιών ανά θέμα, και μια αντίστοιχη κατανομή θεμάτων για το κάθε έγγραφο.

3.3.1.1 Τύποι δεδομένων και ορολογία

Το μοντέλο μπορεί να χρησιμοποιηθεί και σε άλλου τύπου διακριτά δεδομένα εκτός από κείμενα, επιτρέποντας σε σύνολα παρατηρήσεων να εξηγούνται από ομάδες που δεν έχουν παρατηρηθεί και που εξηγούν γιατί ορισμένα τμήματα των δεδομένων είναι παρόμοια. Παρ΄ όλα αυτά, μέσα στη δημοσίευση, οι συγγραφείς

αναφέρονται σε όρους όπως «λέξη», «έγγραφο» και «σώμα εγγράφων», γιατί βοηθάει στη διαίσθηση.

- Μια *λέξη* είναι η βασική μονάδα διακριτών δεδομένων, που ορίζεται ως ένα στοιχείο από ένα λεξιλόγιο με δείκτες $\{1, \ldots, V\}$. Οι λέξεις αναπαριστώνται χρησιμοποιώντας μοναδιαία διανύσματα που έχουν ένα μόνο στοιχείο ίσο με ένα και όλα τα υπόλοιπα στοιχεία ίσα με το μηδέν. Έτσι, χρησιμοποιώντας εκθέτες για τον συμβολισμό των στοιχείων, η v-οστή λέξη στο λεξιλόγιο αντιπροσωπεύεται από έναν V-διάστατο διάνυσμα w όπου $w^v = 1$ και $w^u = 0$ για κάθε $u \neq v$.
- Ένα *έγγραφο* είναι μια ακολουθία N λέξεων που συμβολίζεται ως $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, όπου w_n είναι η n-οστή λέξη στην ακολουθία.
- Ένα *σώμα εγγράφων* είναι μια συλλογή από M έγγραφα που συμβολίζονται ως $\mathcal{D} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M)$.

3.3.1.2 Γενετική διαδικασία των εγγράφων

Το μοντέλο LDA υποθέτει ότι κάθε έγγραφο $\mathbf{w}$ σε ένα σώμα εγγράφων $\mathcal{D}$, το οποίο περιέχει N αριθμό λέξεων, έχει δημιουργηθεί με την ακόλουθη γενετική διαδικασία :

1. Τυχαία επιλογή του $N \sim \text{Poisson}(\xi)$.
2. Τυχαία επιλογή ενός $\theta \sim \text{Dir}(\alpha)$.
3. Για κάθε λέξη w_n από τις N λέξεις του εγγράφου:
   (α′) Τυχαία επιλογή ενός θέματος $z_n \sim \text{Multinomial}(\theta)$.
   (β′) Τυχαία επιλογή μιας λέξης w_n από την $p(w_n|z_n, \beta)$, μια πολυωνυμική (multinomial) πιθανότητα πάνω στο θέμα z_n.

3.3.1.3 Υποθέσεις και παράμετροι

Σε αυτό το βασικό μοντέλο έχουν γίνει αρκετές υποθέσεις. Η διάσταση k της κατανομής Dirichlet θεωρείται γνωστή και σταθερή, επομένως το ίδιο ισχύει και για τη διάσταση της μεταβλητής του θέματος z. Οι πιθανότητες των λέξεων παραμετροποιούνται από έναν πίνακα β διαστάσεων $k \times N$, όπου $\beta_{ij} = p(w^j = 1|z^i = 1)$, οι οποίες προσωρινά θεωρούνται σταθερές, αλλά θα πρέπει να υπολογιστούν. Η υπόθεση ότι το μήκος του εγγράφου ακολουθεί κατανομή Poisson μπορεί να αλλάξει με τη χρήση πιο ρεαλιστικών κατανομών. Ο αριθμός

των λέξεων του εγγράφου N θεωρείται ανεξάρτητος από τις μεταβλητές θ και $\mathbf{z}$, άρα θεωρείται βοηθητική μεταβλητή η οποία συχνά δεν λαμβάνεται υπόψη στην ανάλυση.

Μια k-διάστατη τυχαία μεταβλητή Dirichlet θ έχει την ακόλουθη πυκνότητα πιθανότητας:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} a_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{3.4}$$

όπου α είναι ένα διάνυσμα k διαστάσεων με στοιχεία $\alpha_i > 0$ και $\Gamma(x)$ είναι η συνάρτηση γάμμα.

Δεδομένων των παραμέτρων α και β, η από κοινού κατανομή ενός μείγματος θεμάτων θ, ενός συνόλου $\mathbf{z}$ από N θέματα και ενός συνόλου N λέξεων $\mathbf{w}$ δίνεται από τη σχέση:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.5}$$

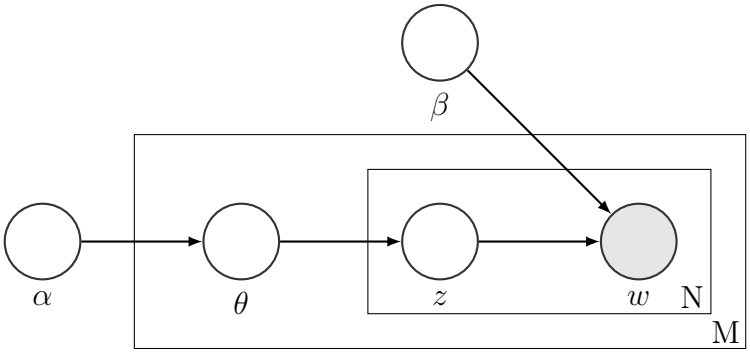
όπου $p(z_n|\theta)$ είναι απλά το θ_i για συγκεκριμένο i τέτοιο ώστε $z_n^i = 1$. Ολοκληρώνοντας ως προς θ και αθροίζοντας ως προς z , προκύπτει η περιθώρια (marginal) κατανομή για ένα έγγραφο:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{3.6}$$

Τελικά, υπολογίζοντας το γινόμενο των πιθανοτήτων των εγγράφων, προκύπτει η πιθανότητα του σώματος εγγράφων:

$$p(\mathcal{D}|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \tag{3.7}$$

Η γραφική αναπαράσταση του μοντέλου LDA απεικονίζεται στο σχήμα 3.1. Όπως φαίνεται και από το σχήμα, υπάρχουν τρία επίπεδα στην αναπαράσταση του LDA. Οι παράμετροι α και β λειτουργούν στο επίπεδο του σώματος εγγράφων, και θεωρείται ότι εκτιμώνται μία φορά, κατά τη διαδικασία δημιουργίας του σώματος. Οι μεταβλητές θ_d, λειτουργούν σε επίπεδο εγγράφου, και εκτιμώνται μία φορά για κάθε έγγραφο. Τέλος, οι μεταβλητές w_{dn} και z_{dn} αφορούν το επίπεδο των λέξεων, και εκτιμώνται μια φορά για κάθε λέξη σε κάθε έγγραφο.



Σχήμα 3.1: Γραφική αναπαράσταση του μοντέλου LDA. Τα ορθογώνια πλαίσια υποδηλώνουν πολλαπλά αντικείμενα. Το εξωτερικό πλαίσιο αναπαριστά τα έγγραφα, ενώ το εσωτερικό πλαίσιο αναπαριστά την επαναλαμβανόμενη επιλογή θεμάτων και λέξεων μέσα σε ένα έγγραφο.

3.3.1.4 Εκτίμηση των παραμέτρων

Η διαδικασία παραγωγής των εγγράφων που περιγράφηκε προηγουμένως, υποθέτει ότι οι κατανομές των θεμάτων, και κατ΄ επέκταση οι παράμετροι του προβλήματος, είναι εξ΄ αρχής γνωστές. Ο βασικός, όμως, στόχος του μοντέλου (και γενικά της θεματικής μοντελοποίησης) είναι η αντίστροφη διαδικασία, δηλαδή ο προσδιορισμός των παραμέτρων του μοντέλου, δεδομένης μιας συλλογής εγγράφων. Η εκ των υστέρων κατανομή των κρυφών μεταβλητών ενός εγγράφου δίνεται από τη σχέση:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{3.8}$$

Δυστυχώς, η συγκεκριμένη κατανομή είναι δύσκολο να υπολογιστεί. Παρ΄ όλα αυτά, υπάρχει μεγάλη ποικιλία προσεγγιστικών αλγορίθμων που μπορούν να χρησιμοποιηθούν για τον LDA. Συνήθως για την προσέγγιση της κατανομής χρησιμοποιείται η δειγματοληψία Gibbs [37].

3.4 Διανυσματικές παραστάσεις λέξεων

Μία από τις ισχυρότερες τάσεις στην Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) τα τελευταία χρόνια είναι η χρήση των διανυσματικών παραστάσεων λέξεων (word embeddings). Πολλές μέθοδοι της Επεξεργασίας Φυσικής Γλώσσας «κωδικοποιούν» λέξεις σε διανύσματα. Τα διανύσματα αυτά, είναι απλά μια αναπαράσταση της κάθε λέξης σε έναν d-διάστατο χώρο. Λόγω της φύσης της ανθρώπινης γλώσσας, οι λέξεις δεν είναι ανεξάρτητες η μια από την άλλη, αλλά συσχετίζονται μεταξύ τους (π.χ. συζεύξεις, χρό-

νοι, συνώνυμα, αντώνυμα). Μια σωστή στρατηγική αναπαράστασης θα πρέπει να προβάλλει αυτές τις σχέσεις στον d-διάστατο χώρο. Επομένως, οι διανυσματικές παραστάσεις λέξεων είναι διανύσματα τα οποία αντιστοιχούν σε λέξεις και έχουν τη βασική ιδιότητα ότι οι σχετικές ομοιότητες τους συσχετίζονται με τη σημασιολογική ομοιότητα των λέξεων στις οποίες αντιστοιχούν.

Τέτοιου τύπου διανύσματα, τα οποία κωδικοποιούν την ομοιότητα ανάμεσα σε όρους, βρίσκουν εφαρμογή σε πολλά προβλήματα της Επεξεργασίας Φυσικής Γλώσσας, όπως είναι η ταξινόμηση κειμένου (text classification), η συσταδοποίηση κειμένων (document clustering), η επισήμανση μερών του λόγου (part of speech tagging), η αναγνώριση ονομάτων οντοτήτων (named entity recognition), η ανάλυση συναισθήματος (sentiment analysis), κ.λπ..

Οι διανυσματικές παραστάσεις λέξεων αποτελούν παράδειγμα επιτυχημένου μοντέλου μη επιβλεπόμενης μάθησης. Το βασικό πλεονέκτημά τους είναι ότι δεν απαιτούν την ύπαρξη επισημάνσεων, αλλά προκύπτουν από μεγάλα σώματα κειμένων με μόνη πληροφορία τη θέση των λέξεων μέσα σε αυτά.

Σε αυτή την ενότητα θα εξηγήσουμε τη θεωρία πίσω από τις διανυσματικές παραστάσεις λέξεων, για ποιο λόγο είναι χρήσιμες στην επεξεργασία φυσικής γλώσσας, και τον τρόπο με τον οποίο προκύπτουν.

3.4.1 Κωδικοποίηση one-hot

Όταν προσπαθούμε να λύσουμε ένα πρόβλημα της Επεξεργασίας Φυσικής Γλώσσας, ο πιο προφανής τρόπος να αναπαραστήσουμε την είσοδο με τη μορφή διανυσμάτων, είναι με τη χρήση της κωδικοποίησης one-hot, όπου κάθε λέξη αντιπροσωπεύεται από ένα διάνυσμα μήκους N, όπου N είναι ο συνολικός αριθμός λέξεων στο λεξιλόγιο. Κάθε δείκτης στα N-διάστατα διανύσματα ανήκει σε μια συγκεκριμένη λέξη, επομένως για κάθε λέξη η αντίστοιχη θέση θα παίρνει την τιμή 1, ενώ οι υπόλοιπες θέσεις θα παίρνουν την τιμή 0.

Για παράδειγμα, η κωδικοποίηση one-hot για την πρόταση 'the cat sat on the mat' θα είναι η εξής:

$$\begin{bmatrix} the \\ cat \\ sat \\ on \\ the \\ mat \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Αυτός ο τρόπος κωδικοποίησης έχει δύο μειονεκτήματα. Το πρώτο είναι ότι όσο αυξάνεται το λεξιλόγιο, αυξάνεται και η διάσταση των διανυσμάτων, τα οποία όμως παραμένουν αραιά. Το δεύτερο μειονέκτημα είναι ότι τα συγκεκριμένα διανύσματα δεν αντιπροσωπεύουν την έννοια των λέξεων.

Για παράδειγμα, αν έχουμε 10000 μοναδικές λέξεις σε ένα σώμα κειμένων, ο πίνακας των one-hot διανυσμάτων που αντιπροσωπεύουν το συγκεκριμένο σώμα κειμένων θα είναι διάστασης $[N_w \times 10000]$, όπου N_w είναι ο συνολικός αριθμός λέξεων. Αντ΄ αυτού μπορούμε να βρούμε ένα σύνολο βαρών, τα βάρη ενσωμάτωσης (embedding weights), τα οποία όταν πολλαπλασιαστούν με τις εισόδους, έχουν ως αποτέλεσμα μια νέα κωδικοποιημένη αναπαράσταση της εισόδου. Αν ο πίνακας βαρών έχει διάσταση $[10000 \times 300]$, μετά από τον πολλαπλασιασμό των δύο πινάκων καταλήγουμε σε μια νέα αναπαράσταση της εισόδου με διαστάσεις $[N_w \times 300]$. Αυτό έχει ως αποτέλεσμα κάθε μοναδική λέξη του σώματος κειμένων πλέον να αντιπροσωπεύεται από ένα διάνυσμα 300 στοιχείων, γεγονός που λύνει το πρόβλημα της αραιότητας των διανυσμάτων.

Στην επόμενη ενότητα θα παρουσιαστεί η βιβλιοθήκη word2vec, η οποία περιέχει δύο μοντέλα τα οποία εκπαιδεύονται με στόχο την εύρεση τέτοιων, χαμηλών διαστάσεων αναπαραστάσεων, οι οποίες ικανοποιούν την αρχική απαίτηση ότι οι θέσεις των διανυσμάτων αναπαράστασης συσχετίζονται με τη σημασιολογία των λέξεων.

3.4.2 Η μεθοδολογία word2vec

Μια από τις πιο αποτελεσματικές στρατηγικές για τη δημιουργία διανυσματικών παραστάσεων λέξεων είναι η βιβλιοθήκη word2vec [64]. Οι κατανεμημένες αναπαραστάσεις λέξεων, όπως είναι το word2vec, λειτουργούν με βάση την υπόθεση της κατανομής (distributional hypothesis) [43, 30], η οποία αναφέρει ότι η έννοια μιας λέξης μπορεί να συναχθεί από τα συμφραζόμενα. Εάν δύο λέξεις μπορούν να καταλάβουν την ίδια θέση σε μια πρόταση, τότε υπάρχει σχέση μεταξύ τους.

Παρ΄ ότι οι περισσότερες μέθοδοι για τις διανυσματικές παραστάσεις λέξεων προέρχονται από την κοινότητα της βαθιάς μηχανικής μάθησης (deep learning), η χρήση βαθιών νευρωνικών δικτύων δεν είναι απαραίτητη για τη δημιουργία καλών διανυσματικών παραστάσεων λέξεων. Τα πιο επιτυχημένα πρόσφατα μοντέλα, το Skipgram και το Continuous Bag-of-Words (CBOW) [65], τα οποία περιέχονται στη βιβλιοθήκη word2vec[1], είναι νευρωνικά δίκτυα με μικρό αριθμό στρωμάτων.

Τα δύο μοντέλα (Skipgram και CBoW) διαφέρουν στον τρόπο με τον οποίο δίνονται η είσοδος και η έξοδος στο νευρωνικό δίκτυο. Κατά τα άλλα, η εκπαίδευση ακολουθεί την ίδια λογική, αφού και στις δύο περιπτώσεις, το word2vec χρησιμοποιεί την ίδια αρχιτεκτονική για να μάθει τις διανυσματικές παραστάσεις των λέξεων με μη επιβλεπόμενο τρόπο. Στις επόμενες παραγράφους περιγράφουμε τις διαφορές των δύο μοντέλων.

[1] https://code.google.com/archive/p/word2vec/

3.4.2.1 Το μοντέλο Skip-gram

Η βασική ιδέα πίσω από το μοντέλο Skip-gram είναι για κάθε λέξη που δίνεται ως είσοδος, το μοντέλο να προβλέπει τις γειτονικές λέξεις (τις λέξεις που υπάρχουν στα συμφραζόμενα). Για παράδειγμα, έστω ότι έχουμε ως δεδομένα την παρακάτω πρόταση:

```
the quick brown fox jumped over the lazy dog
```

Αν για τα συμφραζόμενα ορίσουμε παράθυρο μεγέθους 1, τότε το μοντέλο πρέπει να μπορεί να βρίσκει τις λέξεις που βρίσκονται ακριβώς δεξιά και αριστερά από τη λέξη εισόδου, επομένως τα δεδομένα εκπαίδευσης για την παραπάνω πρόταση θα είναι τα εξής:

```
(quick, the), (quick, brown), (brown, quick), (brown, fox), ...
```

όπου κάθε ζεύγος είναι της μορφής (είσοδος, έξοδος). Το μοντέλο εκπαιδεύεται δίνοντας τη λέξη εξόδου, και προσαρμόζοντας τα βάρη με βάση τη λέξη εξόδου. Η διαδικασία αυτή επαναλαμβάνεται για όλο το σύνολο εκπαίδευσης.

3.4.2.2 Το μοντέλο CBOW

Το μοντέλο continuous bag of words (CBOW) δημιουργεί τα δεδομένα εκπαίδευσης για την προσαρμογή των βαρών με διαφορετικό τρόπο. Αντί να προβλέπει τις γειτονικές λέξεις από την αρχική λέξη, παίρνει ως είσοδο τις γειτονικές λέξεις και προσπαθεί να προβλέψει την αρχική λέξη. Επομένως, τα δεδομένα εκπαίδευσης για την ίδια πρόταση με πριν (με παράθυρο μεγέθους 1) πλέον παίρνουν την εξής μορφή:

```
('the brown', 'quick'), ('quick fox', 'brown'), ...
```

όπου για το πρώτο ζεύγος εισόδου-εξόδου, αθροίζονται τα διανύσματα των λέξεων 'the' και 'brown' και αναμένεται ως έξοδος η λέξη 'quick'.

3.4.2.3 Σύγκριση των δύο μοντέλων

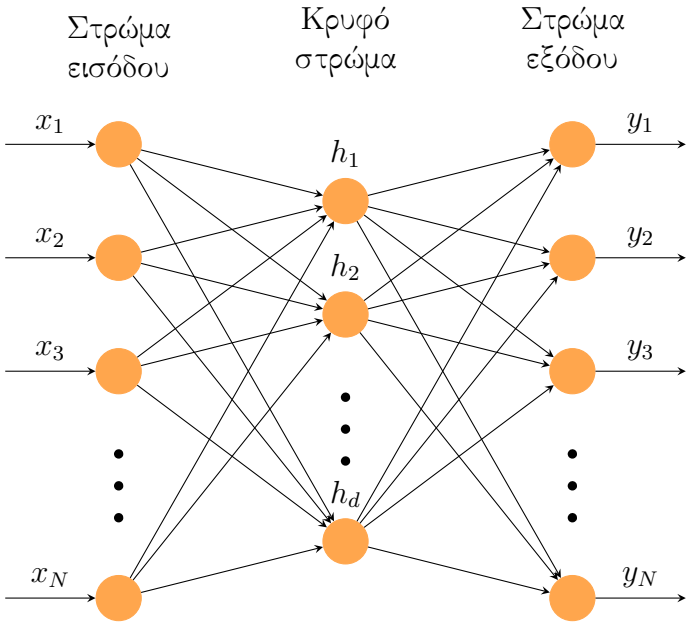
Παρόλο που το CBOW (πρόβλεψη του στόχου από τα συμφραζόμενα) και το Skip-gram (πρόβλεψη των συμφραζομένων από το στόχο) είναι απλώς το ένα ανεστραμμένη μέθοδος του άλλου, κάθε ένα έχει τα πλεονεκτήματα και τα μειονεκτήματά του. Δεδομένου ότι το CBOW μπορεί να χρησιμοποιήσει πολλές λέξεις (τα συμφραζόμενα) για να προβλέψει μια λέξη (τη λέξη-στόχο),

ουσιαστικά εξομαλύνει την κατανομή. Αυτό είναι μιας μορφής κανονικοποίηση, η οποία προσφέρει πολύ καλή απόδοση όταν τα δεδομένα εισόδου δεν είναι μεγάλα. Ωστόσο, το μοντέλο Skip-gram είναι πιο λεπτομερές, επομένως εξάγει περισσότερη πληροφορία και δίνει ακριβέστερες αναπαραστάσεις σε μεγάλα σύνολα δεδομένων.

3.4.2.4 Εκπαίδευση του μοντέλου

Το μοντέλο του νευρωνικού δικτύου που χρησιμοποιείται στο word2vec είναι πολύ απλό στην πιο βασική του μορφή. Η βασική ιδέα είναι η εκπαίδευση ενός νευρωνικού δικτύου, με ένα μοναδικό κρυφό (hidden) στρώμα, το οποίο όμως τελικά δεν θα χρησιμοποιηθεί για το σκοπό που εκπαιδεύτηκε. Ο τελικός στόχος είναι απλά ο υπολογισμός των βαρών του κρυφού στρώματος, τα οποία τελικά θα αποτελέσουν τα διανύσματα των λέξεων τα οποία αναζητούμε. Η ίδια λογική συναντάται και σε άλλα μοντέλα μηχανικής μάθησης (π.χ. στους autoencoders).

Η αρχιτεκτονική του μοντέλου φαίνεται στην εικόνα 3.2, όπου το διάνυσμα $\boldsymbol{x}$ αντιστοιχεί στην αναπαράσταση one-hot της λέξης (ή των λέξεων) εισόδου, ενώ το διάνυσμα $\boldsymbol{y}$ αντιστοιχεί στην αναπαράσταση one-hot της λέξης εξόδου. Τα $\boldsymbol{x}$ και $\boldsymbol{y}$ είναι διαφορετικά για τα δύο μοντέλα, όπως εξηγήθηκε στις παραγράφους 3.4.2.1 και 3.4.2.2.



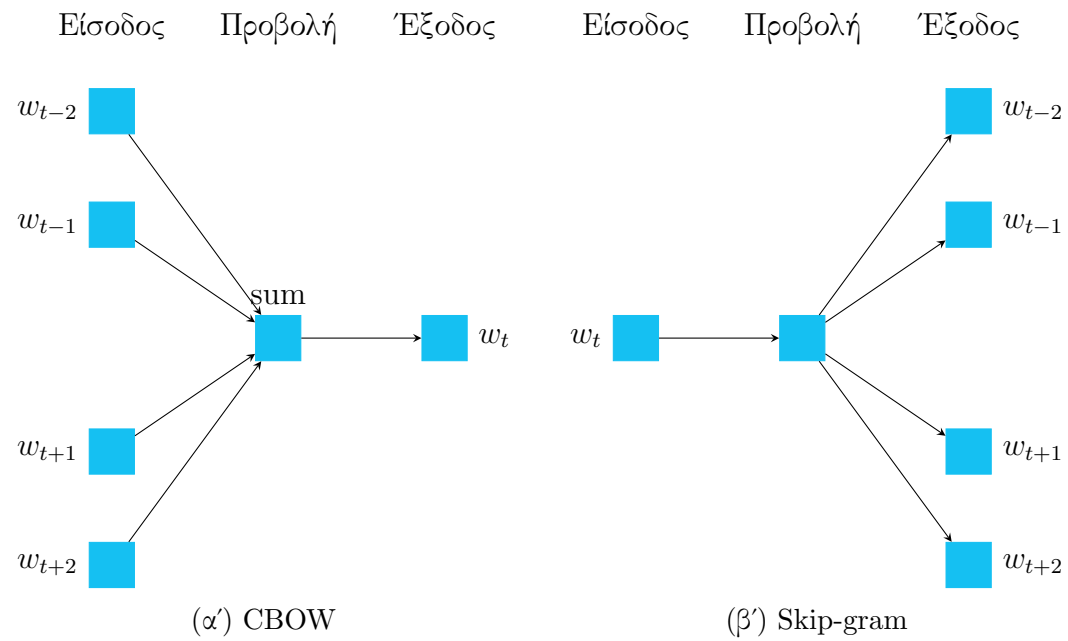
Σχήμα 3.2: Αρχιτεκτονική του μοντέλου word2vec

Και στις δύο περιπτώσεις, όμως, η είσοδος είναι ίση με N, είναι δηλαδή ίση

με το μέγεθος των one-hot διανυσμάτων. Το ίδιο ισχύει και για το στρώμα εξόδου. Τα βάρη του κρυφού στρώματος θα αποτελέσουν τα νέα διανύσματα των λέξεων, επομένως το κρυφό στρώμα έχει d νευρώνες, όπου το d είναι ίσο με το επιθυμητό μέγεθος των διανυσματικών παραστάσεων. Αυτό σημαίνει ότι δεν υπάρχει κάποιος περιορισμός στην επιλογή του μεγέθους, απλά λογικά θα είναι κάποιος αριθμός μικρότερος του N, αφού ένας από τους στόχους είναι η μείωση των διαστάσεων των διανυσμάτων.

Οι ενεργοποιήσεις των κόμβων του κρυφού στρώματος είναι απλά γραμμικές αθροίσεις των σταθμισμένων εισόδων (επομένως δεν εφαρμόζεται κάποια συνάρτηση ενεργοποίησης), ενώ στο στρώμα εξόδου συνήθως χρησιμοποιείται η συνάρτηση softmax. Κατά την εκπαίδευση του δικτύου, τα βάρη ενημερώνονται με τέτοιο τρόπο, ώστε οι γειτονικές λέξεις μιας λέξης εισόδου να δίνουν μεγαλύτερη πιθανότητα στην έξοδο.

Η διαφορά των δύο μοντέλων φαίνεται στην εικόνα 3.3, όπου είναι εμφανής ο τρόπος με τον οποίο δίνεται η είσοδος και η έξοδος. Το CBOW προβλέπει την τρέχουσα λέξη w_t βάσει των συμφραζομένων $w_{t\pm c}$, όπου $c \in \{1, 2, ..., C\}$, ενώ το Skip-gram προβλέπει τις περιβάλλουσες λέξεις $w_{t\pm c}$, δεδομένης μιας λέξης εισόδου w_t.



(α′) CBOW (β′) Skip-gram

Σχήμα 3.3: Οι διαφορές των δύο μοντέλων. Το CBOW προβλέπει την τρέχουσα λέξη βάσει των συμφραζόμενων, ενώ το Skip-gram προβλέπει τις περιβάλλουσες λέξεις με δεδομένη την τρέχουσα λέξη.

3.4.2.5 Επεκτάσεις του μοντέλου word2vec

Το μοντέλο word2vec είναι μια αποτελεσματική μέθοδος για την εκμάθηση διανυσματικών παραστάσεων λέξεων, για τη βελτίωση του οποίου έχουν προταθεί αρκετές επεκτάσεις. Σε αυτή την ενότητα θα αναφερθούμε σε δύο από αυτές, την *Ιεραρχική softmax* [66], η οποία αποτελεί προσέγγιση της συνάρτησης softmax με το πλεονέκτημα ότι έχει μειωμένη υπολογιστική πολυπλοκότητα, και τη *Δειγματοληψία Αρνητικών* (Negative Sampling) [65], μια μέθοδο που προτάθηκε ως επέκταση του Skip-gram, με στόχο τη βελτίωση τόσο της ποιότητας των διανυσμάτων όσο και της ταχύτητας εκπαίδευσης.

3.4.2.5.1 Ιεραρχική softmax

Η συνάρτηση softmax είναι μια συνάρτηση που λαμβάνει ως είσοδο ένα διάνυσμα διαστάσεων K, το οποίο περιέχει πραγματικους αριθμούς, και επιστρέφει μια κατανομή πιθανότητας που αποτελείται από K πιθανότητες, οι οποίες είναι ανάλογες με τα εκθετικά των αριθμών εισόδου. Μετά την εφαρμογή της softmax, κάθε στοιχείο του διανύσματος θα ανήκει στο διάστημα $(0, 1)$, ενώ το άθροισμα των K στοιχείων θα ισούται με 1, έτσι ώστε να μπορούν να ερμηνευτούν ως πιθανότητες. Επιπλέον, τα στοιχεία που έχουν μεγαλύτερες τιμές αντιστοιχίζονται σε μεγαλύτερες πιθανότητες.

Η συνάρτηση softmax ορίζεται ως εξής:

$$softmax(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}, \text{ για } i = 1, \ldots, K \text{ και } \boldsymbol{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K \tag{3.9}$$

Με απλά λόγια, η εκθετική συνάρτηση εφαρμόζεται σε κάθε στοιχείο z_i του διανύσματος εισόδου $\boldsymbol{z}$, και αυτές οι τιμές κανονικοποιούνται διαιρώντας με το άθροισμα όλων των εκθετικών. Αυτό έχει ως αποτέλεσμα ότι το άθροισμα των στοιχείων του διανύσματος εξόδου $softmax(\boldsymbol{z})$ θα είναι ίσο με 1.

Η softmax χρησιμοποιείται συχνά στο στρώμα εξόδου νευρωνικών δικτύων, με στόχο την αντιστοίχιση της μη κανονικοποιημένης εξόδου ενός δικτύου σε μια κατανομή πιθανότητας των κλάσεων εξόδου που προβλέπει το δίκτυο. Η υπολογιστική πολυπλοκότητα του υπολογισμού της συνάρτησης είναι ο αριθμός των λέξεων του λεξιλογίου, $O(N)$.

Στην περίπτωση του word2vec, η πιθανότητα μιας λέξης w δεδομένου ενός πλαισίου c (και κατ΄ επέκταση η αντίστοιχη έξοδος του νευρωνικού) μπορεί να υπολογιστεί με τη βοήθεια της softmax ως εξής:

$$P(w|c) = y = \frac{e^{(h^\top v'_w)}}{\sum_{k=1}^{K} e^{(h^\top v'_{w_k})}} \tag{3.10}$$

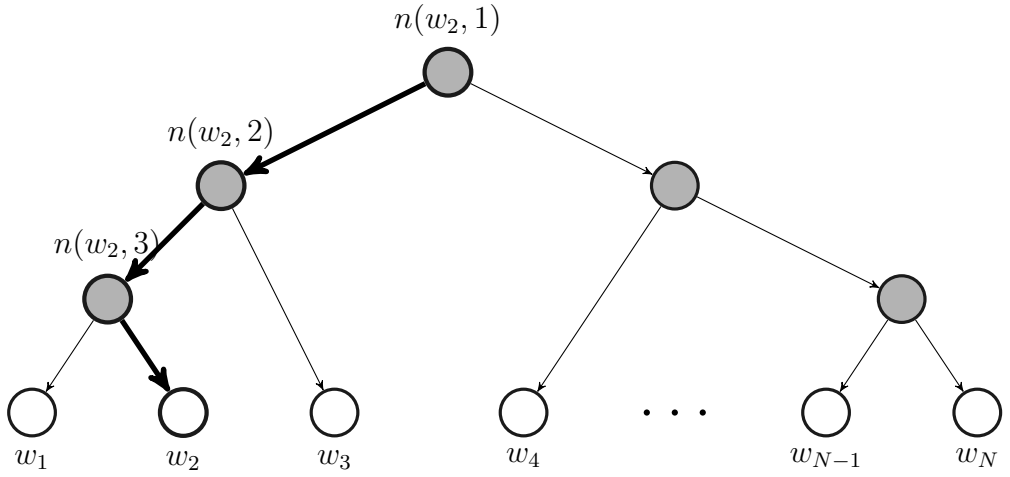
όπου h είναι το διάνυσμα εξόδου του προτελευταίου στρώματος και v'_{w_k} είναι η διανυσματική παράσταση που προκύπτει στην έξοδο για τη λέξη w_k.

Η πολυπλοκότητα της softmax μπορεί να μειωθεί με τη χρήση της ιεραρχικής softmax [66]. Η ιεραρχική softmax κατασκευάζει μια δενδρική δομή για το σώμα κειμένου, όπου κάθε φύλλο του δέντρου αντιπροσωπεύει μια λέξη του λεξιλογίου. Για τον υπολογισμό της πιθανότητας κάθε λέξης απαιτείται η διάσχιση του δέντρου από τη ρίζα μέχρι τη λέξη, και η πιθανότητα της λέξης είναι το γινόμενο της πιθανότητας επιλογής των ακμών που βρίσκονται στη διαδρομή από τη ρίζα στη λέξη. Ένα παράδειγμα τέτοιου δέντρου δίνεται στο σχήμα 3.4.

Μπορούμε να φτάσουμε σε οποιαδήποτε λέξη επιλέγοντας την κατάλληλη διαδρομή από τη ρίζα του δέντρου. Έστω ότι $n(w, j)$ είναι ο j-οστός κόμβος στη διαδρομή από τη ρίζα στη λέξη w, και $L(w)$ είναι το μήκος αυτής της διαδρομής (που σημαίνει ότι $n(w, 1)$ είναι η ρίζα και $n(w, L(w))$ ο κόμβος που αντιστοιχεί στη λέξη). Τότε, η πιθανότητα της λέξης w δίνεται από την εξίσωση:

$$P(w|c) = \prod_{j=1}^{L(w)-1} P(n(w, j) \rightarrow n(w, j+1)|c) \tag{3.11}$$

όπου το σύμβολο c αναφέρεται στα συμφραζόμενα της λέξης. Είναι λογικό ότι οι πιθανότητες μιας λέξης αλλάζουν ανάλογα με το πλαίσιο μέσα στο οποίο βρίσκεται η λέξη.



Σχήμα 3.4: Ιεραρχική softmax.

Με αυτό τον τρόπο, αν το δέντρο δομηθεί σωστά, η πολυπλοκότητα μειώνεται από $O(N)$ σε $O(log(N))$. Το ερώτημα που παραμένει είναι με ποιον τρόπο μπορούν να υπολογιστούν οι πιθανότητες $P(n(w, j) \rightarrow n(w, j+1)|c)$ για κάθε ακμή του δέντρου.

Αυτές οι τιμές παράγονται με τη χρήση της σιγμοειδούς συνάρτησης:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.12}$$

Στη συγκεκριμένη περίπτωση, το x είναι ίσο με το εσωτερικό γινόμενο της διανυσματικής παράστασης της εισόδου h με τη διανυσματική παράσταση της εξόδου του κόμβου n, η οποία συμβολίζεται v'_n:

$$x = h^\top v'_n \tag{3.13}$$

Κατά τη διάσχιση του δέντρου, σε κάθε διακλάδωση πρέπει να μπορεί να υπολογιστεί η πιθανότητα επιλογής της δεξιάς ή της αριστερής ακμής. Για το λόγο αυτό, σε κάθε κόμβο ανατίθεται μια αναπαράσταση. Σε αντίθεση με την κανονική softmax, δεν υπάρχουν πλέον διανυσματικές παραστάσεις v'_w για την έξοδο κάθε λέξης w, αλλά υπάρχουν διανυσματικές παραστάσεις v'_n για κάθε κόμβο n.

Δεδομένου ότι υπάρχουν $N-1$ κόμβοι και ο καθένας διαθέτει μια μοναδική αναπαράσταση, ο αριθμός των παραμέτρων της ιεραρχικής softmax είναι σχεδόν ο ίδιος με της κανονικής softmax. Επομένως, η πιθανότητα επιλογής της δεξιάς ακμής σε έναν κόμβο n, δεδομένου του πλαισίου c, δίνεται από την ακόλουθη εξίσωση:

$$P(right|n, c) = \sigma(h^\top v'_n) \tag{3.14}$$

Επομένως, η πιθανότητα επιλογής της αριστερής ακμής δίνεται ως εξής:

$$P(left|n, c) = 1 - P(right|n, c) = 1 - \sigma(h^\top v'_n) = \sigma(-h^\top v'_n) \tag{3.15}$$

Τελικά, η εξίσωση 3.11 μπορεί να γραφτεί ως εξής:

$$P(w|h) = \prod_{j=1}^{L(w)-1} \sigma([\![n(w, j+1) = ch(n(w, j))]\!] \cdot h^\top v'_{n(w,j)}) \tag{3.16}$$

όπου η $ch(n)$ δίνει το δεξί παιδί του κόμβου n, ενώ η $[\![x]\!]$ επιστρέφει 1 αν το x είναι αληθές, και -1 αν το x είναι ψευδές.

Σε αυτό το σημείο, αξίζει να σημειωθεί ότι, επειδή σε κάθε διακλάδωση το άθροισμα των πιθανοτήτων ισούται με 1, το διάνυσμα εξόδου για τις λέξεις έχει και αυτό άθροισμα ίσο με 1, επομένως δεν χρειάζεται να εφαρμοστεί κανονικοποίηση. Επίσης, μια σημαντική παρατήρηση είναι ότι ο τρόπος κατασκευής του δέντρου (δηλαδή ο τρόπος τοποθέτησης των λέξεων στα φύλλα του δέντρου) μπορεί να επηρεάσει σημαντικά την απόδοση.

3.4.2.5.2 Δειγματοληψία Αρνητικών

Όπως εξηγήθηκε προηγουμένως, το Skip-gram δεν είναι ένα βαθύ νευρωνικό δίκτυο, όμως αποτελείται από μεγάλο αριθμό νευρώνων, μιας και τόσο η είσοδος, όσο και η έξοδος έχουν μέγεθος ίσο με το μέγεθος του εκάστοτε λεξιλογίου. Αυτό σημαίνει ότι σε κάθε βήμα της εκπαίδευσης, το δίκτυο πρέπει να ενημερώνει ένα μεγάλο αριθμό βαρών.

Αντί γι΄ αυτό, η δειγματοληψία αρνητικών επιλέγει να ενημερώνει μόνο τα βάρη που αφορούν συγκεκριμένες λέξεις. Πιο συγκεκριμένα, ενημερώνει τα βάρη που σχετίζονται με την λέξη εξόδου (θετικό δείγμα), ενώ επιλέγει δειγματοληπτικά να ενημερώσει τα βάρη k λέξεων που δεν σχετίζονται με την αρχική λέξη, δηλαδή λέξεων που δεν εμφανίζονται μέσα στο παράθυρο της λέξης εισόδου (αρνητικά δείγματα). Τα αρνητικά αυτά δείγματα $\tilde{w}$ προέρχονται από μια κατανομή θορύβου P_{noise}, η οποία συνήθως περιέχει τυχαίες λέξεις του λεξιλογίου. Έτσι, το μοντέλο μαθαίνει να διακρίνει τα πραγματικά δείγματα από τα «φανταστικά», χωρίς να ενημερώνει τα βάρη όλων των αρνητικών δειγμάτων σε κάθε βήμα της εκπαίδευσης.

Δεδομένης μιας λέξης w_t σε μια πρόταση μήκους T, αρχικά ο στόχος του Skip-gram είναι να μεγιστοποιηθεί η εξής συνάρτηση:

$$J = \sum_{\substack{c \in \{1,\dots,C\}, \\ t \in T}} \log p\left(w_{t \pm c} | w_t\right) \tag{3.17}$$

όπου $w_{t \pm c}$ είναι μια άλλη λέξη στην ίδια πρόταση και $2C$ είναι το *παράθυρο* του μοντέλου (δηλαδή το πόση απόσταση πρέπει να έχει μία λέξη για να θεωρείται ότι ανήκει στα συμφραζόμενα της w_t).

Λαμβάνοντας υπόψη τα προηγούμενα, η αντικειμενική συνάρτηση που πρέπει να μεγιστοποιηθεί είναι η εξής:

$$J = \log Q(y = 1 | w_t, w_c) + k \mathop{\mathbb{E}}_{\tilde{w} \sim P_{noise}} \log Q(y = 0 | \tilde{w}, w_c) \tag{3.18}$$

όπου $Q(y = 1 | w_t, w_c)$ είναι η πιθανότητα δυαδικής λογιστικής παλινδρόμησης (binary logistic regression) του μοντέλου να δει τη λέξη w_t με συμφραζόμενα την w_c. Για τους λόγους που εξηγήθηκαν νωρίτερα, δεν χρησιμοποιείται ολόκληρο το μήκος του λεξιλογίου ως αρνητικά δείγματα, αλλά η προσδοκία προσεγγίζεται με την εξαγωγή k δειγμάτων από την P_{noise}.

3.4.3 Άλλες μεθοδολογίες για την εξαγωγή διανυσματικών παραστάσεων λέξεων

Από την έλευση του word2vec, η χρήση των διανυσματικών παραστάσεων λέξεων έχει συγκεντρώσει μεγάλη προσοχή, με τη εφαρμογή τους σε πολλά διαφορετικά προβλήματα της Επεξεργασίας Φυσικής Γλώσσας. Παρά το γεγονός ότι στη συγκεκριμένη εργασία γίνεται χρήση μόνο του word2vec, τα τελευταία χρόνια έχουν προταθεί και άλλοι αλγόριθμοι για την εξαγωγή διανυσματικών παραστάσεων λέξεων, στους οποίους αξίζει να γίνει σύντομη αναφορά.

Ο GloVe [71] είναι ένας αλγόριθμος μη επιβλεπόμενης μάθησης για τη λήψη διανυσματικών παραστάσεων λέξεων. Σε αντίθεση με το word2vec, το οποίο λαμβάνει υπ΄ όψιν το τοπικό πλαίσιο μιας λέξης, η εκπαίδευσή του γίνεται με βάση τα μη μηδενικά στοιχεία ενός καθολικού πίνακα συν-εμφάνισης λέξεων, ο οποίος καταγράφει το πόσο συχνά οι λέξεις εμφανίζονται μαζί σε ένα δεδομένο σώμα κειμένων. Η δημιουργία του πίνακα απαιτεί ένα μόνο «πέρασμα» ολόκληρου του σώματος κειμένων, κατά το οποίο γίνεται ο υπολογισμός των στατιστικών στοιχείων. Η βασική διαίσθηση πίσω από το μοντέλο είναι ότι οι αναλογίες των πιθανοτήτων συν-εμφανίσεων των λέξεων κωδικοποιούν με κάποιον τρόπο το νόημα των λέξεων. Ο στόχος του μοντέλου είναι να μάθει αναπαραστάσεις των λέξεων, όπου το εσωτερικό γινόμενο των διανυσμάτων να ισούται με το λογάριθμο της πιθανότητας συν-εμφάνισης των λέξεων. Παρ΄ ότι η εκπαίδευση του GloVe είναι πιο γρήγορη από του word2vec, κανένα από τα δύο δεν φαίνεται να υπερέχει του άλλου, αφού τα αποτελέσματά τους εξαρτώνται σε μεγάλο βαθμό από το σύνολο δεδομένων που χρησιμοποιείται κάθε φορά.

Παρά την ευελιξία και την επιτυχία του word2vec και του GloVe στην καταγραφή των σημασιολογικών ιδιοτήτων των λέξεων, τα μοντέλα αδυνατούν να διακρίνουν μεταξύ των διαφορετικών εννοιών μιας λέξης, υπολογίζοντας μία αναπαράσταση για κάθε λέξη, ανεξάρτητα από τα διαφορετικά πλαίσια μέσα στα οποία εμφανίζεται. Προκειμένου να αντιμετωπιστεί αυτό, πολλές προσεγγίσεις έχουν προσπαθήσει να μοντελοποιήσουν τις διαφορετικές έννοιες των λέξεων.

Δύο τέτοιες προσεγγίσεις είναι οι αλγόριθμοι ELMo και BERT. Οι δύο αυτοί αλγόριθμοι κωδικοποιούν το πλαίσιο μιας δεδομένης λέξης, συμπεριλαμβάνοντας στα διανύσματα αναπαράστασης πληροφορία σχετικά με τις προηγούμενες και τις επόμενες λέξεις, ενώ για κάθε εμφάνιση μίας λέξης παράγουν διαφορετική αναπαράσταση.

Ο ELMo (Embeddings from Language Models) [74] είναι ένας αλγόριθμος αναπαράστασης λέξεων ο οποίος μοντελοποιεί τόσο τα σύνθετα χαρακτηριστικά της χρήσης των λέξεων, όπως είναι η σύνταξη και η σημασιολογία, όσο και το πως οι χρήσεις αυτές διαφοροποιούνται ανάλογα με το γλωσσικό πλαίσιο στο οποίο βρίσκεται η κάθε λέξη. Επομένως, ο συγκεκριμένος αλγόριθμος μοντελοποιεί την πολυσημία μιας λέξης. Οι αναπαραστάσεις προκύπτουν από

την εσωτερική κατάσταση ενός αμφίδρομου LSTM [45] (bidirectional LSTM) δύο επιπέδων. Ο ELMo παράγει πολλαπλές αναπαραστάσεις για κάθε λέξη, οι οποίες έχει αποδειχτεί ότι υπερτερούν σε σχέση με τις αναπαραστάσεις που δίνουν τα word2vec και GloVe σε διάφορες εργασίες της Επεξεργασίας Φυσικής Γλώσσας.

Την ίδια λογική ακολουθεί και ο BERT (Bidirectional Encoder Representations from Transformers) [25], ο οποίος κάνει χρήση του μοντέλου μετασχηματιστή (transformer) [107] για τη γλωσσική μοντελοποίηση. Ο μετασχηματιστής αποτελείται από δύο ξεχωριστούς μηχανισμούς, τον κωδικοποιητή (encoder), ο οποίος διαβάζει το κείμενο εισόδου, και τον αποκωδικοποιητή (decoder), ο οποίος παράγει μια πρόβλεψη. Εφόσον ο στόχος είναι η δημιουργία ενός γλωσσικού μοντέλου, ο BERT χρησιμοποιεί μόνο τον κωδικοποιητή. Σε αντίθεση με τα παραδοσιακά μοντέλα, που εκπαιδεύονται ώστε να προβλέπουν την επόμενη λέξη μιας ακολουθίας λέξεων, ο BERT «καλύπτει» κάποιες λέξεις μέσα σε μία πρόταση και προσπαθεί να τις προβλέψει, χρησιμοποιώντας πληροφορία από όλη την πρόταση ταυτόχρονα, ανεξάρτητα από τη θέση των λέξεων. Αυτή η απλή ιδέα, οδηγεί σε κορυφαία αποτελέσματα σε ένα ευρύ φάσμα εργασιών της Επεξεργασίας Φυσικής Γλώσσας.

Κεφάλαιο 4

Ανίχνευση κοινοτήτων στο Twitter

Η ανίχνευση κοινοτήτων (community detection), ή αλλιώς ομαδοποίηση γράφου ή δικτύου (graph/network clustering), είναι ένα από τα πιο δημοφιλή θέματα της σύγχρονης επιστήμης δικτύων, που επιχειρεί να λύσει το πρόβλημα του εντοπισμού της κοινοτικής δομής σε δίκτυα. Τα περισσότερα δίκτυα εμφανίζουν κοινοτική δομή, δηλαδή οι κορυφές τους είναι οργανωμένες σε ομάδες, που ονομάζονται κοινότητες, ομάδες ή συστάδες.

Η ανίχνευση κοινοτήτων δεν είναι ένα σαφώς ορισμένο πρόβλημα, καθώς δεν υπάρχει ένας αυστηρός και καθολικά αποδεκτός ορισμός για το τι είναι κοινότητα. Ο ορισμός αλλάζει ανάλογα με την εφαρμογή, δηλαδή με το ερευνητικό ερώτημα που καλούμαστε κάθε φορά να απαντήσουμε ή το συγκεκριμένο σύστημα το οποίο βρίσκεται υπό μελέτη.

Σαν αποτέλεσμα, δεν υπάρχει σαφής τρόπος αξιολόγησης της απόδοσης των διαφόρων αλγορίθμων, πράγμα που δυσκολεύει και τον τρόπο σύγκρισης των αλγορίθμων μεταξύ τους. Από τη μία, αυτό επιβραδύνει την πρόοδο στην επίλυση ενός ήδη ιδιαίτερα δύσκολου υπολογιστικά προβλήματος, από την άλλη αυτή η ασάφεια αφήνει μεγάλη ελευθερία στις διαφορετικές προσεγγίσεις που προτείνονται για το πρόβλημα.

Σε αυτό το κεφάλαιο θα προσπαθήσουμε να περιγράψουμε το εν λόγω ερευνητικό πρόβλημα, ξεκινώντας με κάποιες βασικές έννοιες της θεωρίας γράφων, οι οποίες είναι απαραίτητες για την κατανόηση των επί μέρους ζητημάτων. Στη συνέχεια, θα δώσουμε τις βασικές απαιτήσεις για την ύπαρξη κοινοτήτων, καθώς και τις τρεις μεγάλες κατηγορίες ορισμών για το τι είναι η κοινότητα, ενώ θα αναφερθούμε στις μεθοδολογίες για την ανίχνευση κοινοτήτων στα κοινωνικά δίκτυα. Τέλος, θα παρουσιάσουμε μια μεθοδολογία για την ανίχνευση κοινοτήτων στο Twitter.

4.1 Βασικές έννοιες θεωρίας γράφων

Κάθε δίκτυο μπορεί να αναπαρασταθεί με τη μορφή γράφου. Γράφος $G = (V, E)$ ονομάζεται ένα σύνολο κορυφών ή κόμβων V και ένα σύνολο ακμών $E \subseteq V \times V$, οι οποίες συνδέουν ζεύγη κόμβων. Ένας γράφος μπορεί να είναι κατευθυνόμενος (directed) ή μη κατευθυνόμενος (undirected), ανάλογα με το αν οι ακμές αποτελούν διατεταγμένα ζεύγη κορυφών, ή όχι. Σε πολλά πραγματικά παραδείγματα, οι ακμές ενός γράφου περιέχουν βάρη, οπότε αποκαλείται σταθμισμένος (weighted).

Ο αριθμός κόμβων στο γράφο είναι ίσος με $n = |V|$ και ο αριθμός των ακμών με $m = |E|$. Το μέγιστο μέγεθος ενός γράφου ισούται με το συνολικό αριθμό μη διατεταγμένων ζευγών κόμβων, δηλαδή με $n(n-1)/2$. Αν $|E| = n(n-1)/2$, ο γράφος αποτελεί κλίκα (clique) ή πλήρη (complete) γράφο, και συμβολίζεται με K_n.

Όλη η πληροφορία σχετικά με την τοπολογία ενός γράφου περιέχεται στον πίνακα γειτνίασης A, ο οποίος είναι ένας πίνακας διαστάσεων $n \times n$, του οποίου το στοιχείο A_{ij} ισούται με 1 αν υπάρχει ακμή που συνδέει τις κορυφές i και j, διαφορετικά είναι ίσο με μηδέν. Για ένα μη κατευθυνόμενο γράφο, ο πίνακας A είναι συμμετρικός. Αν οι ακμές έχουν βάρη, τότε ορίζεται ο πίνακας βαρών W, του οποίου το στοιχείο W_{ij} εκφράζει το βάρος της ακμής μεταξύ των κορυφών i και j.

Εάν δύο κόμβοι συνδέονται με μία ακμή, τότε ονομάζονται γειτονικοί. Για ένα μη κατευθυνόμενο γράφο, ο βαθμός (degree) ενός κόμβου είναι ο αριθμός των προσκείμενων σε αυτόν ακμών. Ο βαθμός ενός κόμβου v συμβολίζεται με deg(v) ή deg v.

4.2 Ορισμός της κοινότητας

Όπως αναφέραμε και προηγουμένως, δεν υπάρχει ένας μοναδικός ορισμός για την κοινότητα. Ένας διαισθητικός τρόπος να ορίσουμε τις κοινότητες είναι ως ομάδες κόμβων οι οποίες είναι πιο πυκνά συνδεδεμένες μεταξύ τους παρά με το υπόλοιπο δίκτυο.

Έστω ότι C είναι ένας υπογράφος του γράφου G, όπου το πλήθος κόμβων είναι ίσο με n_c. Ορίζουμε τον εσωτερικό (internal) και τον εξωτερικό βαθμό (external degree) ενός κόμβου $v \in C$, k_v^{int} και k_v^{ext}, ως το πλήθος των ακμών που συνδέουν τον v με άλλους κόμβους του υπογράφου C ή με τον υπόλοιπο γράφο αντίστοιχα. Στην περίπτωση που $k_v^{ext} = 0$, ο κόμβος v έχει γείτονες μόνο μέσα στον C, που σημαίνει ότι ο C πιθανώς αποτελεί μια καλή ομάδα για τον v. Αντίθετα, αν $k_v^{int} = 0$, τότε ο κόμβος v δεν έχει καμία σύνδεση με τον C, επομένως θα έπρεπε να ανατεθεί σε άλλη ομάδα.

Ορίζουμε την ενδο-συσταδική πυκνότητα (intra-cluster density) $\delta_{int}(C)$ του υπογράφου C ως το λόγο του πλήθους των εσωτερικών ακμών του C προς το πλήθος όλων των δυνατών εσωτερικών ακμών, δηλαδή:

$$\delta_{int}(C) = \frac{\#\text{ εσωτερικών ακμών του } C}{n_c(n_c-1)/2} \tag{4.1}$$

Αντίστοιχα, η δια-συσταδική πυκνότητα (inter-cluster density) $\delta_{ext}(C)$ είναι ο λόγος του πλήθους των ακμών που συνδέουν τους κόμβους του C με τον υπόλοιπο γράφο προς το πλήθος όλων των δυνατών ακμών, δηλαδή:

$$\delta_{ext}(C) = \frac{\#\text{ δια-ομαδικών ακμών του } C}{n_c(n-n_c)/2} \tag{4.2}$$

Σύμφωνα με τα παραπάνω, προκειμένου ο C να αποτελεί κοινότητα, η ενδο-συσταδική πυκνότητα $\delta_{int}(C)$ αναμένεται να είναι αισθητά μεγαλύτερη από τη μέση πυκνότητα ακμών $\delta(G)$ του G, η οποία δίνεται από το λόγο του πλήθους των ακμών του G προς το πλήθος όλων των δυνατών ακμών $n(n-1)/2$. Από την άλλη, η δια-συσταδική πυκνότητα $\delta_{ext}(C)$ αναμένεται να είναι πολύ μικρότερη από τη $\delta_{int}(G)$. Η αναζήτηση του καλύτερου συμβιβασμού μεταξύ μιας μεγάλης $\delta_{int}(G)$ και μιας μικρής $\delta_{ext}(C)$ είναι άμεσα ή έμμεσα ο στόχος κάθε αλγορίθμου ανίχνευσης κοινοτήτων.

Τα παραπάνω συνιστούν τις βασικές απαιτήσεις για την κατανόηση του τι είναι η κοινότητα. Με βάση αυτά, μπορούμε να διακρίνουμε τρεις κατηγορίες ορισμών [31]: τοπικούς, καθολικούς και βασισμένους στην ομοιότητα των κόμβων.

Τοπικοί ορισμοί: Οι τοπικοί ορισμοί επικεντρώνονται σε έναν υπογράφο, συμπεριλαμβανομένης, πιθανώς, και της άμεσης γειτονιάς του, τον οποίο και μελετούν, αγνοώντας το υπόλοιπο δίκτυο. Η λογική πίσω από αυτό είναι ότι οι κοινότητες έχουν λίγους δεσμούς με το υπόλοιπο δίκτυο, άρα, σε κάποιο βαθμό, μπορούν να θεωρηθούν ως ξεχωριστές οντότητες με δική τους αυτονομία. Οι κοινότητες που προκύπτουν είναι οι μέγιστοι υπογράφοι, στους οποίους δεν μπορούν να προστεθούν νέοι κόμβοι και ακμές χωρίς να χάσουν την ιδιότητα που τους ορίζει. Υπάρχουν τέσσερις τύποι κριτηρίων: η πλήρης αμοιβαιότητα (complete mutuality), η προσβασιμότητα (reachability), ο βαθμός των κορυφών (vertex degree) και η σύγκριση μεταξύ εσωτερικής και εξωτερικής συνοχής (comparison of internal versus external cohesion)

Καθολικοί ορισμοί: Οι κοινότητες μπορούν επίσης να οριστούν σε σχέση με ολόκληρο το γράφο ως σύνολο. Αυτό είναι λογικό στην περίπτωση που οι ομάδες κόμβων αποτελούν ουσιώδη μέρη του γράφου, τα οποία δεν μπορούν

να διαχωριστούν χωρίς να επηρεάσουν τη λειτουργία του συστήματος. Έχουν προταθεί πολλά καθολικά κριτήρια για την ανίχνευση των κοινοτήτων. Στις περισσότερες περιπτώσεις είναι έμμεσοι ορισμοί, στους οποίους μια καθολική ιδιότητα του γράφου χρησιμοποιείται σε έναν αλγόριθμο ο οποίος τελικά παρέχει τις κοινότητες. Ωστόσο, υπάρχει μια κλάση ορισμών που βασίζονται στην ιδέα ότι ένας γράφος έχει κοινοτική δομή εάν διαφέρει σημαντικά από έναν τυχαίο γράφο (κατά Erdös-Rényi - οποιεσδήποτε δύο κορυφές του γράφου έχουν την ίδια πιθανότητα να είναι συνδεδεμένες).

Ορισμοί βασισμένοι στην ομοιότητα των κόμβων: Οι εν λόγω ορισμοί βασίζονται στην υπόθεση ότι οι κοινότητες είναι ομάδες κόμβων παρόμοιων μεταξύ τους. Η ομοιότητα μεταξύ κάθε ζεύγους κόμβων μπορεί να υπολογιστεί με βάση κάποια ιδιότητα, τοπική ή καθολική, ανεξάρτητα από το αν υπάρχει ακμή που συνδέει τους κόμβους ή όχι. Κάθε κόμβος καταλήγει στην ομάδα με τους πιο όμοιους με αυτόν κόμβους. Η χρήση μέτρων ομοιότητας για τη συσταδοποίηση αποτελεί τη βάση των παραδοσιακών μεθόδων, όπως είναι η ιεραρχική, η μερική και η φασματική συσταδοποίηση.

4.3 Ανίχνευση κοινοτήτων στα κοινωνικά δίκτυα

Τα τελευταία χρόνια, η δημοτικότητα των κοινωνικών δικτύων έχει αυξηθεί δραματικά. Παρ΄ ότι τα περισσότερα κοινωνικά δίκτυα έχουν δημιουργηθεί σχετικά πρόσφατα (στα τελευταία δεκαπέντε χρόνια), ένα ευρύ φάσμα επιστημονικών ερευνών και μεθόδων έχει δημοσιευτεί για την κατανόηση και την αποκάλυψη της υποκείμενης δομής αυτών των πολύπλοκων δικτύων, με πληθώρα εφαρμογών σε πολλά πεδία. Πιο συγκεκριμένα, η αποκάλυψη των υποκείμενων κοινοτήτων σε δίκτυα του πραγματικού κόσμου μπορεί να ρίξει φως στις δομικές ιδιότητες και στον τρόπο λειτουργίας τους.

4.3.1 Αναπαράσταση των κοινωνικών δικτύων μέσω γράφων

Στην περίπτωση των κοινωνικών δικτύων το πρόβλημα γίνεται πιο πολύπλοκο. Πλέον, έχουμε ένα ευρύ φάσμα αντικειμένων που συνδέονται μεταξύ τους μέσω διαφορετικών τύπων αλληλεπιδράσεων και σχέσεων.

Αρχικά, οι κόμβοι μπορεί να είναι διαφορετικού τύπου, είναι πιθανό για παράδειγμα να αναπαριστούν χρήστες, περιεχόμενο (κείμενο, εικόνες, βίντεο), ακόμα και μεταδεδομένα (θεματικές κατηγορίες ή επισημάνσεις). Επιπρόσθετα,

οι ακμές μπορεί να είναι πολλών διαφορετικών τύπων, ενώ μπορεί να έχουν βάρη (ή όχι) και να έχουν κατεύθυνση (ή όχι), αναλόγως με τη φύση του κοινωνικού δικτύου που μελετάται. Και σε αυτή την περίπτωση το δίκτυο μπορεί να αναπαρασταθεί με τη βοήθεια ενός γράφου $G = (V, E)$, όμως μέσα στα V και E υπάρχουν υποσύνολα κόμβων και ακμών, αντίστοιχα, τα οποία αποτελούν αντικείμενα του ίδιου τύπου [70]. Στην περίπτωση που ο γράφος είναι κατευθυνόμενος, το πρόβλημα γίνεται πιο δύσκολο, αφού πλέον οι πίνακες γειτνίασης και βαρών δεν είναι συμμετρικοί, που σημαίνει ότι ο ορισμός του προβλήματος αλλάζει, ενώ απαιτούνται πιο σύνθετες τεχνικές για την επίλυσή του [63].

Σε κάποιες περιπτώσεις, οι σχέσεις ανάμεσα στους κόμβους μπορούν να αναπαρασταθούν με τη μορφή ενός πολυεπίπεδου γράφου, ο οποίος αποτελείται από πολλούς ανεξάρτητους γράφους, όπου κάθε γράφος αντιπροσωπεύει μια πτυχή των σχέσεων [52].

4.3.2 Προσεγγίσεις της ανίχνευσης κοινοτήτων στα κοινωνικά δίκτυα

Οι προσεγγίσεις για την ανίχνευση κοινοτήτων στα κοινωνικά δίκτυα μπορούν να χωριστούν σε δύο κατηγορίες [26]: στις προσεγγίσεις που βασίζονται στην τοπολογία (topology-based approaches) και στις προσεγγίσεις που βασίζονται στη θεματολογία (topic-based approaches).

Παρ΄ όλα αυτά, οι κοινότητες που ανιχνεύονται λαμβάνοντας υπ΄ όψιν μόνο την τοπολογία του δικτύου τείνουν να περιλαμβάνουν διαφορετικές θεματικές ενότητες, ενώ κάθε θεματική κοινότητα (που έχει προκύψει από τις προσεγγίσεις που βασίζονται στη θεματολογία) μπορεί να χωριστεί σε αξιόλογες υποκοινότητες βασισμένες στην τοπολογία. Επομένως, η ανίχνευση κοινοτήτων θα πρέπει να εξετάζει τόσο τη δομή του δικτύου, όσο και την κειμενική πληροφορία.

Επομένως, πολλές διαφορετικές προσεγγίσεις συνδυάζουν την τοπολογία των κοινωνικών συνδέσεων και τα θεματικά χαρακτηριστικά προκειμένου να εξάγουν σημαντικά θέματα συζήτησης μεταξύ χρηστών [57, 88, 59]. Οι συγγραφείς του [116] χρησιμοποιούν έναν αλγόριθμο ομαδοποίησης για να χωρίσουν τα μέλη του συνόλου δεδομένων τους σε θεματικές συστάδες και στη συνέχεια πραγματοποιούν ανάλυση των συνδέσεων σε κάθε συστάδα για να ανιχνεύσουν τις κοινότητες, ενώ στο [111], τόσο η θεματική ομοιότητα μεταξύ των χρηστών, όσο και οι συνδέσεις, λαμβάνονται υπ΄ όψιν για τη μέτρηση της επιρροής των χρηστών του Twitter.

Στο [115], οι συγγραφείς ανακαλύπτουν κοινότητες στο Twitter με βάση τα ενδιαφέροντα των χρηστών, εφαρμόζοντας την έννοια της ομοιότητας των χρηστών, τόσο από το περιεχόμενο κειμένου, όσο και την κοινωνική δομή, χρησι-

μοποιώντας αλγόριθμους κλασικής ομαδοποίησης. Στο [34], παρουσιάζεται ένα πλαίσιο βασισμένο στην μηχανική μάθηση, το οποίο είναι ικανό να ανακαλύπτει τους κορυφαίους παρόμοιους χρήστες για κάθε χρήστη του Twitter, με βάση την ομοιότητα του περιεχομένου που παράγει ο χρήστης.

Η διαδικασία της αναγνώρισης του συνόλου των κόμβων που έχουν τη μεγαλύτερη επιρροή σε έναν κόμβο για ένα δεδομένο θέμα εισάγεται στο [61], η οποία, τελικά, έχει ως αποτέλεσμα την εύρεση των κόμβων που αντιπροσωπεύουν την πηγή για κάθε θέμα. Στο [72], ο χαρακτηρισμός και η κατηγοριοποίηση ιστολογίων και άλλων σύντομων κειμένων βελτιώνεται με την ανάλυση και τον εντοπισμό των χαρακτηριστικών που μπορούν να τα διακρίνουν. Στο [78], οι συγγραφείς προτείνουν ένα μοντέλο το οποίο χρησιμοποιεί δύο μοντέλα λανθάνουσας ανάθεσης Dirichlet (LDA) για την ομαδοποίηση παρόμοιων τηλεοπτικών χρηστών και παρόμοιων περιγραφών τηλεοπτικών προγραμμάτων ταυτόχρονα.

Οι συγγραφείς στο [85] επεξεργάζονται και αναλύουν το περιεχόμενο που δημοσιεύτηκε στις κοινωνικές υπηρεσίες για να εκτιμήσουν αυτόματα από το κείμενο τις προσωπικότητες των ανθρώπων και να δώσουν πολύτιμες συστάσεις σε άτομα με παρόμοιους τύπους προσωπικότητας. Άλλες κατευθύνσεις περιλαμβάνουν την ανίχνευση επικαλυπτόμενων κοινοτήτων [58, 113] ή ενοποιητικές προσεγγίσεις όπως το μοντέλο στο [114] που χρησιμοποιεί τόσο την δομή των ακμών του γράφου όσο και τα χαρακτηριστικά των κόμβων.

4.4 Περιγραφή της μεθοδολογίας

Το Twitter[1] είναι μια διαδικτυακή υπηρεσία ειδήσεων και κοινωνικής δικτύωσης, η οποία δημιουργήθηκε το Μάρτιο του 2006 από τους Jack Dorsey, Noah Glass, Biz Stone και Evan Williams. Από το 2016, το Twitter είχε περισσότερους από 319 εκατομμύρια ενεργούς χρήστες μηνιαίως.

Το Twitter επιτρέπει στους χρήστες του να επικοινωνούν και να μοιράζονται τις σκέψεις και τις δραστηριότητές τους μέσω σύντομων μηνυμάτων, τα οποία ονομάζονται *tweets*. Επιπλέον, οι χρήστες του Twitter μπορούν να εγγραφούν στα tweets άλλων χρηστών «ακολουθώντας» τους ("follow"). Αυτό μπορεί να αποτελεί ένδειξη των ενδιαφερόντων του χρήστη, καθώς οι χρήστες τείνουν να ακολουθούν λογαριασμούς με βάση το πόσο ενδιαφέροντα θεωρούν τα tweets τους. Ως εκ τούτου, η ανάλυση των σχέσεων και του περιεχομένου που μοιράζονται οι χρήστες μπορεί να δώσει πληροφορίες για τα ενδιαφέροντά τους, τις απόψεις και τη συμπεριφορά τους.

Σε αυτή την ενότητα θα παρουσιάσουμε μια μεθοδολογία για την ανίχνευση κοινοτήτων στο Twitter. Όπως αναφέρθηκε προηγουμένως, ο ορισμός της

[1] https://twitter.com/

κοινότητας αλλάζει ανάλογα με την εφαρμογή. Στη συγκεκριμένη εργασία, ορίζουμε τις κοινότητες ως ομάδες χρηστών που είναι πιο πυκνά συνδεδεμένες μεταξύ τους παρά με το υπόλοιπο δίκτυο, αλληλεπιδρούν περισσότερο μεταξύ τους και μοιράζονται κοινά ενδιαφέροντα.

4.4.1 Ομαδοποίηση των χρηστών σε κοινότητες

Η προτεινόμενη μεθοδολογία αποτελείται από δύο βήματα. Αρχικά, ορίζουμε την έννοια της ομοιότητας (similarity) ανάμεσα στους χρήστες του Twitter και υπολογίζουμε την απόσταση για κάθε ζεύγος χρηστών. Το δεύτερο βήμα χρησιμοποιεί τις υπολογισμένες αποστάσεις με στόχο τη συσταδοποίηση (clustering) των χρηστών σε κοινότητες.

4.4.1.1 Ομοιότητα χρηστών

Η ομοιότητα ανάμεσα σε ένα ζεύγος χρηστών του Twitter προκύπτει από τις μεταξύ τους αλληλεπιδράσεις, όπως έχουν καταγραφεί στο ιστορικό των tweets τους. Επομένως, η ομοιότητα μπορεί να υπολογιστεί με βάση όλα τα χαρακτηριστικά στοιχεία του Twitter που παρέχουν πληροφορία για τις αλληλεπιδράσεις ενός χρήστη με τους υπόλοιπους χρήστες: τη λίστα των φίλων (friends) και των ακολούθων (followers) του, τα hashtags που περιλαμβάνονται στα tweets του, τις απαντήσεις (replies) του σε άλλους χρήστες και τους χρήστες που αναφέρονται με άλλο τρόπο μέσα στα tweets του χρήστη (user mentions). Οι μετρικές ομοιότητας που βασίζονται σε αυτά τα χαρακτηριστικά στοιχεία του Twitter θα παρουσιαστούν στις επόμενες παραγράφους.

4.4.1.1.1 Ομοιότητα με βάση τη σχέση ακολούθησης (following relationship) Ο πιο προφανής τρόπος για να καθορίσουμε την ομοιότητα ανάμεσα σε δύο χρήστες είναι να εξετάσουμε κατά πόσο ο ένας ακολουθεί τον άλλο. Είναι γνωστό ότι ένας χρήστης του Twitter μπορεί να ακολουθεί έναν άλλο χρήστη χωρίς να ισχύει το ανάποδο, επομένως ο γράφος είναι κατευθυνόμενος. Έστω u_i και u_j δύο χρήστες του Twitter. Η ομοιότητά τους με βάση τη σχέση ακολούθησης υπολογίζεται ως εξής:

$$S_1(u_i, u_j) = \begin{cases} 1, & \text{αν ο χρήστης } u_i \text{ ακολουθεί τον } u_j \\ & \text{και ο } u_j \text{ ακολουθεί τον } u_i \\ 0.5, & \text{αν μόνο ένας από τους χρήστες } u_i,\ u_j \\ & \text{ακολουθεί τον άλλο} \\ 0, & \text{αλλιώς} \end{cases} \tag{4.3}$$

4.4.1.1.2 Ομοιότητα με βάση τους κοινούς ακόλουθους (common followers) Η σελίδα κάθε χρήστη στο Twitter εμφανίζει μια ροή που αποτελείται από τα tweets των λογαριασμών που έχει επιλέξει να ακολουθήσει ο χρήστης, που σημαίνει ότι η επιλογή των λογαριασμών που ακολουθεί κάποιος είναι ένδειξη των ενδιαφερόντων του. Επομένως, οι χρήστες που ενδιαφέρονται για παρόμοια θέματα αναμένεται να έχουν έναν αριθμό κοινών ακολούθων. Συνεπώς, μπορούμε να υπολογίσουμε την ομοιότητα χρηστών με βάση τους κοινούς ακόλουθους όπως φαίνεται στην εξίσωση 4.4, όπου $followers_i$ είναι το σύνολο των ακολούθων του χρήστη u_i και n είναι ο αριθμός χρηστών. Κανονικοποιούμε τον αριθμό των κοινών ακολούθων, ώστε να διασφαλίσουμε ότι όλες οι τιμές κυμαίνονται μεταξύ 0 και 1.

$$S_2(u_i, u_j) = \frac{|followers_i \cap followers_j|}{\max_{1 \leq l \leq n} followers_l} \tag{4.4}$$

4.4.1.1.3 Ομοιότητα με βάση τους κοινούς φίλους (common friends) Στο Twitter, το σύνολο των χρηστών που ακολουθεί κάποιος ονομάζονται φίλοι (friends). Όπως αναφέρθηκε προηγουμένως, μπορούμε να υποθέσουμε ότι παρόμοιοι χρήστες θα έχουν κοινούς φίλους. Λαμβάνοντας υπ΄ όψιν τα παραπάνω, ορίζουμε τη ομοιότητα των χρηστών ως εξής:

$$S_3(u_i, u_j) = \frac{|friends_i \cap friends_j|}{\max_{1 \leq l \leq n} friends_l} \tag{4.5}$$

4.4.1.1.4 Ομοιότητα με βάση τα hashtags Μια λέξη κλειδί ή μία φράση χωρίς κενά που ξεκινάει με το σύμβολο # ονομάζεται hashtag. Σκοπός των hashtags είναι να κατηγοριοποιήσουν τα tweets και, κατ΄ επέκταση, να απλοποιήσουν τον τρόπο με τον οποίο οι χρήστες αναζητούν σχετικά tweets. Για να υπολογίσουμε την ομοιότητα με βάση τα hashtags, πρέπει πρώτα να προσδιορίσουμε τη σημασία κάθε λέξης - κλειδιού για έναν συγκεκριμένο χρήστη.

Αυτό μπορεί να επιτευχθεί υπολογίζοντας τα βάρη tf-idf του μοντέλου διανυσματικού χώρου (vector space model) [91], θεωρώντας ότι όλα τα hashtags που περιέχονται στα tweets ενός χρήστη αποτελούν ένα ενιαίο έγγραφο. Επομένως, εάν h_i είναι το διάνυσμα tf-idf των hashtags που χρησιμοποιούνται από το χρήστη u_i και $cos(x, y)$ είναι η συνάρτηση συνημιτόνου που υπολογίζει την ομοιότητα μεταξύ δύο διανυσμάτων x και y, η ομοιότητα με βάση τα hashtags ορίζεται ως εξής:

$$S_4(u_i, u_j) = cos(h_i, h_j) \tag{4.6}$$

4.4.1.1.5 Ομοιότητα με βάση τις απαντήσεις (replies) Ένας χρήστης του Twitter μπορεί να συμμετάσχει σε μια συνομιλία απαντώντας σε tweets άλλων χρηστών. Η συχνότητα των απαντήσεων μεταξύ δύο χρηστών, καθώς και ο αριθμός των χρηστών στους οποίους και οι δύο χρήστες απαντούν, είναι δύο δείκτες της ομοιότητας των χρηστών, η οποία για τις απαντήσεις ορίζεται στην εξίσωση 4.7. Αν R_i και R_j είναι τα σύνολα των χρηστών στους οποίους έχουν απαντήσει οι χρήστες u_i και u_j, αντίστοιχα, τότε ο αριθμός των χρηστών στους οποίους έχουν απαντήσει και οι δύο είναι $|R_i \cap R_j|$. nr_{ij} είναι ο αριθμός των φορών που ο u_i απάντησε στον u_j, και NR_i είναι ο αριθμός των φορών που ο u_i απάντησε σε tweet άλλου χρήστη.

$$S_5(u_i, u_j) = \frac{|R_i \cap R_j|}{\sqrt{|R_i|}\sqrt{|R_j|}} + \frac{nr_{ij} + nr_{ji}}{NR_i + NR_j} \tag{4.7}$$

4.4.1.1.6 Ομοιότητα με βάση τις αναφορές (user mentions) Τοποθετώντας το σύμβολο @ μπροστά από ένα όνομα χρήστη (username), ένας χρήστης μπορεί να επισημάνει (tag) έναν άλλο χρήστη σε ένα tweet. Ως αναφορά θεωρείται οποιοδήποτε tweet περιέχει στο κείμενο του το "@username", πράγμα που σημαίνει ότι οι απαντήσεις θεωρούνται επίσης αναφορές. Η ομοιότητα χρηστών με βάση τις αναφορές υπολογίζεται ως εξής:

$$S_6(u_i, u_j) = \frac{|M_i \cap M_j|}{\sqrt{|M_i|}\sqrt{|M_j|}} + \frac{nm_{ij} + nm_{ji}}{NM_i + NM_j} \tag{4.8}$$

όπου M_i και M_j είναι τα σύνολα των χρηστών που έχουν αναφέρει οι u_i και u_j στα tweets τους, αντίστοιχα, nm_{ij} είναι ο αριθμός των αναφορών του χρήστη u_i στον u_j, και NM_i είναι ο συνολικός αριθμός αναφορών που έκανε ο u_i.

Οι τελευταίες δύο μετρικές έχουν προσαρμοστεί με βάση μια σχετική μετρική η οποία προτάθηκε από τους Zhang et al. [115].

4.4.1.1.7 Συνολική ομοιότητα Προκειμένου να υπολογιστεί η συνολική ομοιότητα των χρηστών, οι μετρικές αυτές πρέπει να συνδυαστούν. Στην προσέγγισή μας υιοθετήσαμε ένα γραμμικό συνδυασμό των διαφορετικών μετρικών ομοιότητας, όπως φαίνεται στην ακόλουθη εξίσωση:

$$S(u_i, u_j) = \sum_{m=1}^{6} a_m S_m(u_i, u_j) \tag{4.9}$$

Οι τιμές που αντιστοιχούν στις παραμέτρους a_m είναι μεταξύ 0 και 1 και $a_1 + a_2 + a_3 + a_4 + a_5 + a_6 = 1$. Η διαδικασία με την οποία υπολογίζονται οι εν λόγω τιμές περιγράφεται στην ενότητα 5.3.

4.4.1.2 Ομαδοποίηση των χρηστών

Το τελικό βήμα της ανίχνευσης των κοινοτήτων αποτελείται από τη διαδικασία της ομαδοποίησης των χρηστών για το σχηματισμό των κοινοτήτων, λαμβάνοντας υπ΄ όψιν τις μετρικές ομοιότητας για κάθε ζεύγος χρηστών. Όπως είναι γνωστό, υπάρχει ένα ευρύ φάσμα αλγόριθμων συσταδοποίησης (clustering algorithms). Εφόσον οι απόλυτες θέσεις των σημείων για το συγκεκριμένο πρόβλημα δεν είναι διαθέσιμες, ο αλγόριθμος ο οποίος θα επιλεγεί θα πρέπει να παίρνει σαν είσοδο την απόσταση ανάμεσα στα σημεία.

Για αυτόν το λόγο, επιλέχθηκε ο αλγόριθμος Διάδοσης Συνάφειας, ο οποίος, όπως εξηγήθηκε στην ενότητα 2.3.0.1, λαμβάνει ως είσοδο μια συλλογή από ομοιότητες μεταξύ σημείων δεδομένων, ενώ ο αριθμός των συστάδων που προκύπτουν δεν είναι προκαθορισμένος

Σε αυτό το σημείο θα μπορούσε να χρησιμοποιηθεί οποιοσδήποτε αλγόριθμος συσταδοποίησης, υπό την προϋπόθεση ότι δέχεται ως είσοδο έναν πίνακα ομοιοτήτων (ή αποστάσεων) ανάμεσα σε σημεία, και όχι τις θέσεις τους. Επομένως, θα μπορούσε να εφαρμοστεί ιεραρχική συσταδοποίηση [60], σε συνδυασμό με κάποιο κριτήριο για την εξαγωγή των τελικών συστάδων, αφού οι ιεραρχικοί αλγόριθμοι δίνουν σαν έξοδο ένα δενδρόγραμμα, καθώς και παραδοσιακές μέθοδοι ανίχνευσης κοινοτήτων, όπως είναι η φασματική συσταδοποίηση (spectral clustering) [109] ή οι μέθοδοι βελτιστοποίησης της τμηματικότητας (modularity optimization), π.χ. ο αλγόριθμος των Girvan και Newman [33] ή η μέθοδος Louvain [10].

Μια άλλη ενδιαφέρουσα τεχνική είναι ο αλγόριθμος Label Propagation [117, 81]. Ο συγκεκριμένος αλγόριθμος ανιχνεύει την κοινοτική δομή σε δίκτυα, αναθέτοντας επισημάνσεις (labels) στους κόμβους, τις οποίες ενημερώνει σε κάθε επανάληψη, επιλέγοντας για κάθε κόμβο την επισήμανση που εμφανίζεται με τη μεγαλύτερη συχνότητα στους γείτονές του. Παρ΄ όλο που ο συγκεκριμένος αλγόριθμος έχει μικρότερη πολυπλοκότητα σε σύγκριση με αντίστοιχους αλγορίθμους, έχει το μειονέκτημα ότι δεν παράγει μια μοναδική λύση, αλλά ένα σύνολο πολλών λύσεων, ακόμα και για την ίδια αρχική κατάσταση. Ο αριθμός των λύσεων αυτών μπορεί να μειωθεί, αν ένα μικρό, συνήθως, ποσοστό των κόμβων έχουν ήδη επισημάνσεις. Λόγω της φύσης των δεδομένων που μελετάμε, τα οποία είναι πραγματικά, μη επισημασμένα δεδομένα, δεν είναι δυνατή η εφαρμογή του συγκεκριμένου αλγόριθμου.

4.4.2 Εξαγωγή των θεμάτων και προσθήκη επισημάνσεων

Στην τρέχουσα ενότητα, μελετάμε το περιεχόμενο που μοιράζονται οι χρήστες για να καθορίσουμε αν οι χρήστες που ανήκουν στην ίδια συστάδα τείνουν

να συζητούν στα tweets τους για τα ίδια θέματα. Συγκεκριμένα, περιγράφουμε τη διαδικασία της εξαγωγής των θεμάτων που συζητήθηκαν από τους χρήστες, προτείνουμε μια μέθοδο αφαίρεσης των θεμάτων που δεν παρουσιάζουν ενδιαφέρον και περιγράφουμε μια διαδικασία για την αυτόματη παραγωγή επισημάνσεων για κάθε θέμα με τη χρήση της λίστας των λέξεων-κλειδιών. Επιπρόσθετα, ο συνδυασμός της διαδικασίας εξαγωγής θεμάτων και της διαδικασίας δημιουργίας επισημάνσεων μπορεί να θεωρηθεί ως μια νέα μεθοδολογία ανίχνευσης της θεματολογίας, δεδομένου ότι οι επισημάνσεις που δημιουργούνται δεν περιλαμβάνονται κατ΄ ανάγκη στα tweets των χρηστών, αλλά δημιουργούνται με τη χρήση της Wikipedia, όπως θα δούμε στην ενότητα 4.4.2.3.

4.4.2.1 Εξαγωγή των θεμάτων με τη χρήση της μεθόδου LDA

Σε ένα βήμα προεπεξεργασίας, οντότητες όπως είναι τα hashtags, οι αναφορές σε χρήστες και οι διευθύνσεις URL αφαιρούνται από το κείμενο κάθε tweet του συνόλου δεδομένων. Ακολούθως, το κείμενο χωρίζεται σε λέξεις, αφαιρούνται οι λέξεις με μικρή διακριτική ικανότητα (stopwords) και για τις λέξεις γίνεται απαλοιφή των καταλήξεων (stemming). Τέλος, οι λέξεις που περιέχονται στα tweets του ίδιου χρήστη συγκεντρώνονται σε ένα έγγραφο. Αυτή η διαδικασία επαναλαμβάνεται για όλους τους χρήστες, επομένως, μετά το πέρας αυτής της διαδικασίας, μπορούμε να θεωρήσουμε όλο το σύνολο δεδομένων μας ως μια συλλογή εγγράφων.

Η εξαγωγή των θεμάτων μπορεί να επιτευχθεί μέσω του μοντέλου λανθάνουσας κατανομής Dirichlet, ένα γενετικό πιθανοτικό μοντέλο ενός σώματος εγγράφων, το οποίο βασίζεται στην ιδέα ότι κάθε έγγραφο είναι ένα μίγμα από θεματικές ενότητες, όπου κάθε θεματική ενότητα αποτελείται από λέξεις με μια τυχαία κατανομή πιθανότητας.

Με τη χρήση του LDA στη συλλογή κειμένων των χρηστών λαμβάνουμε ένα σύνολο N θεμάτων (και την κατανομή των λέξεων ή κλειδιών ανά θέμα) και μια αντίστοιχη κατανομή θεμάτων για το έγγραφο κάθε χρήστη, την *Κατανομή Θεμάτων Χρήστη (User Topic Distribution - UTD)*:

$$UTD_i = [topic_{1,u_i}, topic_{2,u_i}, ..., topic_{N,u_i}] \tag{4.10}$$

όπου $topic_{p,u_i}$ είναι η τιμή της πιθανότητας του θέματος p στην κατανομή των θεμάτων του χρήστη u_i.

Στόχος μας είναι να βρούμε την κατανομή θεμάτων για κάθε ομάδα χρηστών, και κατ΄ επέκταση τα ενδιαφέροντα των χρηστών μέσα στην ομάδα. Επομένως, μπορούμε να θεωρήσουμε ότι όλοι οι χρήστες μιας ομάδας σχηματίζουν ένα έγγραφο, να ενώσουμε τα έγγραφα κάθε ομάδας και να καταλήξουμε σε μια νέα συλλογή εγγράφων όπου κάθε έγγραφο αντιστοιχεί και σε μία ομάδα.

Υπολογίζοντας την κατανομή θεμάτων αυτής της συλλογής ομάδων, διατηρώντας όμως ταυτόχρονα τα ίδια θέματα (την ίδια κατανομή λέξεων - κλειδιών ανά θέμα) με τη συλλογή των χρηστών, καταλήγουμε στην κατανομή των θεμάτων για κάθε ομάδα C_r, την οποία ονομάζουμε *Τοπική Κατανομή Θεμάτων (Local Topic Distribution - LTD)*:

$$LTD_r = [topic_{1,C_r}, topic_{2,C_r}, ..., topic_{N,C_r}] \tag{4.11}$$

όπου $topic_{p,C_r}$ είναι η τιμή της πιθανότητας του θέματος p στο κείμενο που αντιστοιχεί στην ομάδα C_r.

Με αντίστοιχο τρόπο μπορούμε να υπολογίσουμε την συνολική κατανομή θεμάτων, ή *Κατανομή Θεμάτων Συλλογής (Collection Topic Distribution - CTD)*, ενώνοντας όλα τα κείμενα:

$$CTD = [topic_{1,C}, topic_{2,C}, ..., topic_{N,C}] \tag{4.12}$$

όπου $topic_{p,C}$ είναι η τιμή της πιθανότητας του θέματος p για όλη τη συλλογή $C = \bigcup_{r=1}^{k} C_r$ και k είναι ο αριθμός των ομάδων.

Έτσι, μπορούμε να ορίσουμε ένα νέο μέτρο, το *Ενδιαφέρον Τοπικής Θεματολογίας (Local Topic Interestingness - LTI)*, για την ομάδα r:

$$LTI_r = \|CTD - LTD_r\| \tag{4.13}$$

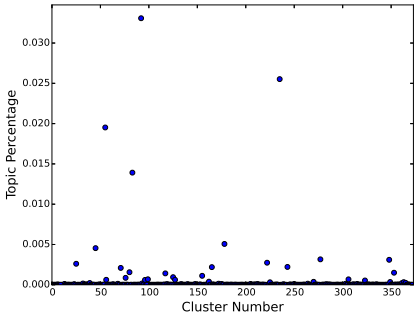
Αυτό το μέτρο παρέχει μια εκτίμηση για το πόσο αποκλίνει η συζήτηση σε μια κοινότητα από τη γενική συζήτηση, άρα και επικεντρώνεται σε πιο ειδικά θέματα.

4.4.2.2 Αφαίρεση των μη αντιπροσωπευτικών θεμάτων

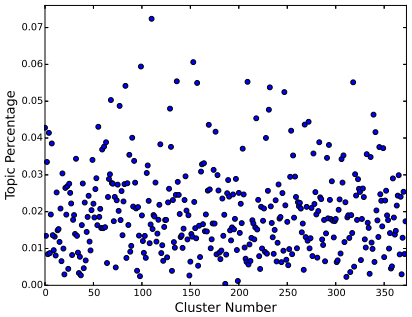
Τα θέματα που εξάγονται με τη χρήση του LDA δεν είναι όλα εξίσου σημαντικά, από την άποψη ότι δεν μας δίνουν πληροφορία σε σχέση με τα ενδιαφέροντα των χρηστών. Ορισμένα θέματα αποτελούνται από γενικές, καθημερινές λέξεις, ενώ άλλα αντιπροσωπεύουν κοινά ενδιαφέροντα για την πλειονότητα των χρηστών.

Για παράδειγμα, λέξεις - κλειδιά όπως "autism, spectrum, support, asperger", "paper, research, student, phd" και "race, vettel, alonso, hamilton" αντιπροσωπεύουν ουσιαστικά και σημαντικά θέματα, ενώ το θέμα "year, week, time, day, today" περιέχει γενικές λέξεις. Επιπλέον, το θέμα "build, develop, base, create, release", το οποίο φαίνεται να είναι αντιπροσωπευτικό των ενδιαφερόντων των χρηστών, δεν θεωρείται αντιπροσωπευτικό σε μια συλλογή εγγράφων τα οποία είναι σχετικά με τον προγραμματισμό.

Ως εκ τούτου, αναπτύξαμε μια νέα μέθοδο για την εξάλειψη των μη αντιπροσωπευτικών θεμάτων. Η μέθοδος αποτελείται από δύο διαφορετικές προσεγγίσεις, που σημαίνει ότι για κάθε θέμα υπολογίζονται δύο διαφορετικά μέτρα,



Σχήμα 4.1: Ποσοστά του θέματος "race, vettel, alonso" για κάθε ομάδα.



Σχήμα 4.2: Ποσοστά του θέματος "year, week, time" για κάθε ομάδα.

τα οποία στη συνέχεια συνδυάζονται με στόχο να ταξινομηθούν τα θέματα με βάση το ενδιαφέρον τους.

4.4.2.2.1 Μέσο ποσοστό ανά ομάδα Το πρώτο από τα δύο μέτρα που εισάγουμε βασίζεται στην ιδέα ότι τα ενδιαφέροντα θέματα συνήθως συζητούνται σε μικρό αριθμό κοινοτήτων. Επομένως, για κάθε ομάδα κανονικοποιούμε τις κατανομές των θεμάτων και υπολογίζουμε το ποσοστό κάθε θέματος για κάθε ομάδα.

Μια γραφική παράσταση των ποσοστών ανά ομάδα για το ενδιαφέρον θέμα "race, vettel, alonso" απεικονίζεται στο σχήμα 4.1. Το ποσοστό αυτού του θέματος είναι πολύ υψηλό για ελάχιστες ομάδες, ενώ είναι σχεδόν μηδενικό για τις υπόλοιπες ομάδες. Αυτό έχει ως αποτέλεσμα το μέσο ποσοστό ανά ομάδα (Mean Cluster Percentage - MCP) να είναι επίσης χαμηλό, κοντά στο μηδέν.

Αντίθετα, για ένα γενικό θέμα όπως το "year, week, time", σημειώνουμε υψηλά ποσοστά σε όλες σχεδόν τις ομάδες (Σχήμα 4.2), με αποτέλεσμα να έχουμε υψηλή τιμή του MCP. Επομένως, γίνεται σαφές ότι οι υψηλές τιμές MCP αντιστοιχούν σε μη αντιπροσωπευτικά θέματα και οι χαμηλές τιμές σε ενδιαφέροντα θέματα.

4.4.2.2.2 Μέση συχνότητα λέξης Κατά την εκτέλεση του αλγορίθμου LDA, λαμβάνουμε για κάθε θέμα μια λίστα που αποτελείται από τις m πιο σημαντικές λέξεις. Η σημασία του θέματος μπορεί να εξαχθεί εξετάζοντας κατά πόσο υπάρχει συνάφεια ανάμεσα στις λέξεις. Αυτό μπορεί να επιτευχθεί με τον υπολογισμό της συχνότητας εμφάνισης κάθε λέξης σε ένα σώμα κειμένων της αγγλικής γλώσσας.

Σε αυτή την εργασία, χρησιμοποιήσαμε το τελευταίο Wikipedia dump[2] για να ανακτήσουμε τη συχνότητα κάθε λέξης. Έτσι, υπολογίζουμε τη μέση συχνότητα λέξεων (Mean Word Frequency - MWF) για τις λέξεις που ανήκουν στη λίστα κάθε θέματος. Αναμένεται, λοιπόν, ότι τα μη αντιπροσωπευτικά θέματα θα έχουν υψηλότερες τιμές MWF σε σύγκριση με τα ενδιαφέροντα.

4.4.2.2.3 Σύνθετος δείκτης κατάταξης Για να συνδυάσουμε τα δύο μέτρα και να ταξινομήσουμε τα θέματα, υπολογίζουμε απλώς τον αριθμητικό μέσο όρο των κανονικοποιημένων τιμών των MCP και MWF. Το νέο αυτό μέτρο το ονομάζουμε σύνθετο δείκτη κατάταξης (Composite Ranking Index - CRI). Στην ενότητα 5.5 παρουσιάζουμε την μέθοδο για τον υπολογισμό του ορίου για τις τιμές του δείκτη, πάνω από το οποίο τα θέματα θεωρούνται μη αντιπροσωπευτικά και κάτω από το οποίο θεωρούνται ενδιαφέροντα.

4.4.2.3 Αυτόματη εξαγωγή επισημάνσεων

Όπως εξηγήθηκε προηγουμένως, κάθε θέμα αντιπροσωπεύεται από μια λίστα λέξεων - κλειδιών. Συνήθως, απαιτείται περαιτέρω επεξεργασία σε αυτή τη λίστα για να αποκαλυφθεί η σημασιολογία ενός θέματος. Για το λόγο αυτό, προτείνουμε μια μεθοδολογία για την αυτόματη δημιουργία επισημάνσεων χρησιμοποιώντας περιεχόμενο που ανακτάται από την αγγλική Wikipedia.

Η πρόσβαση στα απαιτούμενα δεδομένα από τη Wikipedia παρέχεται με τη βοήθεια του MediaWiki API[3], μια υπηρεσία web που παρέχει εύκολη πρόσβαση σε χαρακτηριστικά, δεδομένα και μεταδεδομένα της Wikipedia. Για κάθε λέξη - κλειδί ενός θέματος, αναζητούμε τις πιο σχετικές σελίδες της Wikipedia. Για να προσδιορίσουμε ποια από τις σελίδες αυτές είναι η πιο αντιπροσωπευτική για τη λέξη - κλειδί, υπολογίζουμε τον αριθμό εμφάνισης όλων των λέξεων - κλειδιών του θέματος στο κείμενο της σελίδας. Επιλέγεται η σελίδα με την υψηλότερη τιμή. Αυτή η διαδικασία έχει ως αποτέλεσμα μια μοναδική σελίδα ανά λέξη - κλειδί.

Στη συνέχεια, για κάθε μία από τις παραπάνω σελίδες ανακτούμε τη λίστα των κατηγοριών τους με τη βοήθεια του MediaWiki API. Όλες οι κατηγορίες συγκεντρώνονται και ταξινομούνται με βάση τον αριθμό των σελίδων στις οποίες αντιστοιχούν. Οι πιο συνηθισμένες κατηγορίες επιλέγονται ως επισημάνσεις για το θέμα. Το τελικό αποτέλεσμα είναι ένα μικρό σύνολο λέξεων ή σύντομων φράσεων, αντί για μια μεγάλη λίστα λέξεων - κλειδιών.

[2] `dumps.wikimedia.org`

[3] `www.mediawiki.org/wiki/API:Main_page`

Κεφάλαιο 5

Ανίχνευση κοινοτήτων στο Twitter - Πειραματικό μέρος

Σε αυτό το κεφάλαιο θα παρουσιάσουμε ένα σύνολο πειραμάτων με στόχο την αξιολόγηση της μεθοδολογίας που προτάθηκε για την ανίχνευση κοινοτήτων στο Twitter, όπως αυτή παρουσιάστηκε στο κεφάλαιο 4.

5.1 Σύνολο δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε συλλέχθηκε χρησιμοποιώντας το Twitter Searching API. Δεδομένου ότι στόχος μας ήταν να συμπεριλάβουμε χρήστες με κοινά ενδιαφέροντα στο σύνολο δεδομένων (επομένως χρήστες που κάνουν tweet για παρόμοια θέματα), επιλέξαμε τους ακόλουθους του @isocpp ως τη συλλογή χρηστών μας. Το @isocpp είναι ο λογαριασμός Twitter της επιτροπής προτύπων (standards committee) της ISO C++, οπότε αναμένεται ότι οι χρήστες που ακολουθούν αυτόν τον λογαριασμό θα ενδιαφέρονται σε κάποιο βαθμό και θα κάνουν tweet για τον προγραμματισμό.

Η επιλογή αυτή είχε ως αποτέλεσμα ένα σύνολο 5077 χρηστών. Για καθέναν από αυτούς τους χρήστες συγκεντρώσαμε όλα τα δημοσιευμένα tweets τους, μια λίστα με τα αναγνωριστικά των ακολούθων τους και μια λίστα με τα αναγνωριστικά των φίλων τους.

Από αυτό το αρχικό σύνολο χρηστών, αποκλείσαμε όσους είχαν λιγότερα από είκοσι δημοσιευμένα tweets. Το τελικό σύνολο αποτελείται από 2728 χρήστες. Θα πρέπει επίσης να σημειωθεί ότι ένα από τα μέτρα ομοιότητας που περιγράφηκαν προηγουμένως, το μέτρο ομοιότητας με βάση τη σχέση ακολούθησης (following relationship), έχει μηδενική τιμή για κάθε ζεύγος χρηστών στο σύνολο δεδομένων.

5.2 Κριτήριο ποιότητας συσταδοποίησης

Σε αυτή την ενότητα θα παρουσιαστεί ένα κριτήριο αξιολόγησης της συσταδοποίησης, το οποίο στη συνέχεια θα χρησιμοποιηθεί για την εύρεση των τιμών των παραμέτρων του μοντέλου μας. Οι τυπικές συναρτήσεις για την αξιολόγηση συσταδοποίησης εστιάζουν σε δύο σημεία, στην υψηλή ομοιότητα στο εσωτερικό των συστάδων και τη χαμηλή ομοιότητα ανάμεσα στις συστάδες. Επομένως, στη συγκεκριμένη περίπτωση, ο στόχος είναι να υπάρχουν παρόμοιες συνθέσεις θεμάτων για τους χρήστες μέσα σε μία ομάδα και διαφορετικές συνθέσεις θεμάτων μεταξύ διαφορετικών ομάδων χρηστών. Μια αντικειμενική συνάρτηση που απεικονίζει την ομοιότητα στο εσωτερικό των συστάδων είναι το άθροισμα των τετραγώνων εντός των ομάδων (within-group sum of squares - WGSS) [17]:

$$WGSS = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{u_i \in C_r} \sum_{u_j \in C_r} \|UTD_i - UTD_j\|^2 \tag{5.1}$$

όπου UTD_i και UTD_j είναι οι κατανομές θεμάτων των χρηστών u_i και u_j, όπως περιγράφηκαν στην ενότητα 4.4.2.1, C_r είναι η συστάδα με δείκτη r και n_r είναι ο αριθμός των χρηστών στη συστάδα C_r.

Με αντίστοιχο τρόπο, η ομοιότητα μεταξύ των συστάδων μπορεί να αναπαρασταθεί από το άθροισμα των τετραγώνων μεταξύ των ομάδων (between-group sum of squares - BGSS), το οποίο παρουσιάζεται στην εξίσωση 5.2. Το BGSS απαιτεί τη χρήση των αποστάσεων μεταξύ των κέντρων των συστάδων, αλλά οι απόλυτες θέσεις τους δεν είναι γνωστές. Επομένως, ορίζουμε την απόσταση μεταξύ δύο συστάδων ως την ευκλείδεια απόσταση των τοπικών κατανομών τους.

$$BGSS = \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} \|LTD_r - LTD_s\|^2 \tag{5.2}$$

Ο δείκτης Calinski-Harabasz [17], ο οποίος ονομάζεται και κριτήριο αναλογίας διακύμανσης (variance ratio criterion - VRC), είναι ένα κριτήριο που εφαρμόζεται στην ανάλυση συσταδοποίησης που συνδυάζει τις συναρτήσεις WGSS και BGSS. Ο δείκτης Calinski-Harabasz ορίζεται ως εξής:

$$VRC = \frac{BGSS}{k-1} / \frac{WGSS}{n-k} \tag{5.3}$$

όπου n είναι ο αριθμός των χρηστών και k είναι ο αριθμός των συστάδων. Κατά συνέπεια, η μέγιστη τιμή του δείκτη αντιστοιχεί στη βέλτιστη συσταδοποίση των χρηστών.

5.3 Παράμετροι μετρικών ομοιότητας

Σε αυτή την ενότητα, περιγράφουμε τη διαδικασία υπολογισμού των τιμών των έξι παραμέτρων που εισάγονται στην εξίσωση 4.9. Στόχος μας είναι να καθορίσουμε τον τρόπο με τον οποίο τα μέτρα ατομικής ομοιότητας επηρεάζουν τη συνολική ομοιότητα. Αρχικά, η συνολική ομοιότητα υπολογίζεται για όλους τους πιθανούς συνδυασμούς των παραμέτρων, όπου κάθε παράμετρος παίρνει τιμές από 0 έως 1, με βήμα 0.1, κρατώντας το άθροισμά τους ίσο με 1. Στη συνέχεια, ο αλγόριθμος διάδοσης συνάφειας εκτελείται με κάθε διαφορετική μετρική ομοιότητας ως είσοδο. Το αποτέλεσμα αυτής της διαδικασίας είναι ένας αριθμός πιθανών κατατμήσεων ή ομαδοποιήσεων των χρηστών, οι οποίες πρέπει να αξιολογηθούν.

Ο δείκτης VRC εφαρμόζεται σε όλες τις ομαδοποιήσεις που παράγονται από κάθε εκτέλεση του αλγορίθμου διάδοσης συνάφειας. Οι συνδυασμοί των παραμέτρων που καταλήγουν σε μια κανονικοποιημένη τιμή του δείκτη που είναι μικρότερη από 0.9 απορρίπτονται. Η μέση τιμή για κάθε παράμετρο υπολογίζεται για τους υπόλοιπους συνδυασμούς. Αυτή η διαδικασία έχει ως αποτέλεσμα τις βέλτιστες τιμές οι οποίες παρατίθενται στον πίνακα 5.1.

Πίνακας 5.1: Βέλτιστες τιμές των παραμέτρων a_m

Παράμετρος	Μέτρο ομοιότητας	Τιμή παραμέτρου
a_2	κοινοί ακόλουθοι	0.15
a_3	κοινοί φίλοι	0.1187
a_4	hastags	0.4187
a_5	απαντήσεις	0.1562
a_6	αναφορές χρηστών	0.1562

Παρατηρούμε ότι η παράμετρος που αντιστοιχεί στα hastags έχει μεγαλύτερη τιμή από τις υπόλοιπες παραμέτρους. Αυτό είναι αναμενόμενο, αν λάβουμε υπ΄ όψιν ότι οι βέλτιστες τιμές υπολογίστηκαν χρησιμοποιώντας τις κατανομές των θεμάτων του κειμένου των tweets, ενώ τα hastags αποτελούν επισημάνσεις για τα tweets με συγκεκριμένα θέματα, άρα είναι δείκτες των θεμάτων που συζητούνται σε ένα tweet.

Επιπροσθέτως, εξετάζουμε τον τρόπο με τον οποίο η επιλογή του αριθμού των θεμάτων N επηρεάζει τις τιμές των παραμέτρων a_m. Για το λόγο αυτό, η διαδικασία που περιγράφηκε παραπάνω επαναλαμβάνεται για διαφορετικές τιμές του N. Οι τιμές που προκύπτουν για την καλύτερη ομαδοποίηση και οι μέσες τιμές για τις ομαδοποιήσεις με τιμή VRC μεγαλύτερη από 0.9 παρουσιάζονται στον πίνακα 5.2. Είναι προφανές ότι οι τιμές για την καλύτερη ομαδοποίηση είναι αμετάβλητες όταν ο αριθμός των θεμάτων είναι σχετικά μικρός, ενώ οι μέσες τιμές δεν μεταβάλλονται σημαντικά για τις διαφορετικές τιμές του N.

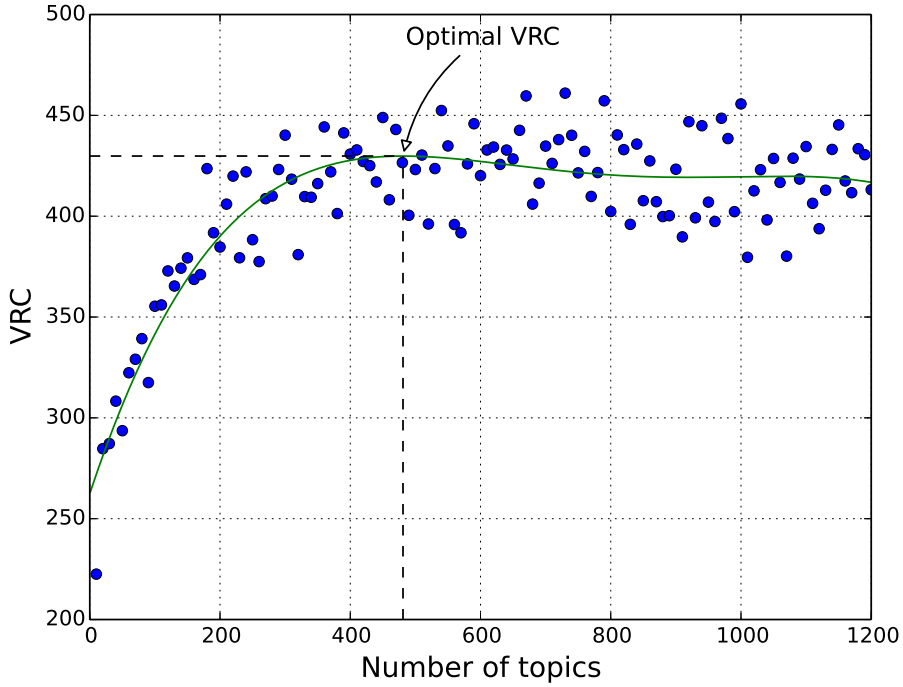
Πίνακας 5.2: Τιμές των παραμέτρων a_m για διαφορετικό αριθμό θεμάτων N

	Βέλτιστη συσταδοποίηση					Μέση τιμή για $VRC > 0.9$				
N	a_2	a_3	a_4	a_5	a_6	a_2	a_3	a_4	a_5	a_6
100	0.1	0.1	0.5	0.1	0.2	0.1500	0.1187	0.4187	0.1562	0.1562
200	0.1	0.1	0.5	0.1	0.2	0.1571	0.1071	0.4214	0.1571	0.1571
300	0.1	0.1	0.5	0.1	0.2	0.1666	0.1166	0.4055	0.1555	0.1555
400	0.1	0.1	0.5	0.1	0.2	0.1666	0.1166	0.4055	0.1555	0.1555
500	0.1	0.1	0.5	0.1	0.2	0.1666	0.1166	0.4055	0.1555	0.1555
600	0.1	0.1	0.5	0.1	0.2	0.1631	0.1263	0.4052	0.1526	0.1526
700	0.1	0.1	0.5	0.1	0.2	0.1588	0.1176	0.4117	0.1588	0.1529
800	0.1	0.1	0.5	0.1	0.2	0.1666	0.1166	0.4055	0.1555	0.1555
900	0.1	0.1	0.5	0.1	0.2	0.1470	0.1294	0.4176	0.1529	0.1529
1000	0.1	0.1	0.6	0.1	0.1	0.1588	0.1176	0.4117	0.1588	0.1529
1100	0.1	0.1	0.6	0.1	0.1	0.1500	0.1187	0.4187	0.1562	0.1562
1200	0.1	0.1	0.6	0.1	0.1	0.1555	0.1277	0.4111	0.1555	0.1500

5.4 Αριθμός θεμάτων

Ένα από τα πιο σημαντικά ζητήματα σχετικά με τον αλγόριθμο LDA είναι η επιλογή του βέλτιστου αριθμού θεμάτων. Ο λόγος είναι ότι ένας μικρός αριθμός θεμάτων παρέχει πιο γενικά θέματα, ενώ ένας μεγάλος αριθμός δυσκολεύει την κατανόηση των ζητημάτων. Η απάντηση εξαρτάται γενικά από το μέγεθος της συλλογής των εγγράφων και το επιθυμητό αποτέλεσμα.

Ο δείκτης Calinski-Harabasz μπορεί να χρησιμοποιηθεί για τον προσδιορισμό του ιδανικού αριθμού θεμάτων για το σύνολο δεδομένων μας. Επομένως, υπολογίζουμε το δείκτη VRC για την ομαδοποίηση που προκύπτει από τις βέλτιστες τιμές των παραμέτρων a_m και για διαφορετικούς αριθμούς θεμάτων. Το αποτέλεσμα απεικονίζεται στο σχήμα 5.1, όπου παρατηρούμε ότι η βέλτιστη τιμή για το VRC αντιστοιχεί στα 450 θέματα.



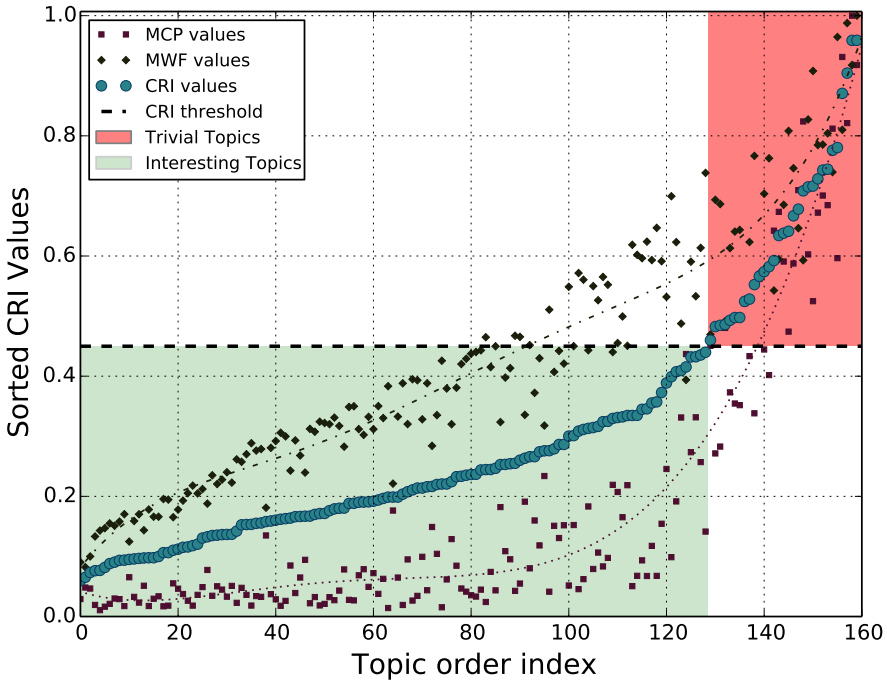
Σχήμα 5.1: Η τιμή του VRC για διαφορετικό αριθμό θεμάτων N.

5.5 Αφαίρεση των μη αντιπροσωπευτικών θεμάτων

Όπως αναφέρθηκε στην ενότητα 4.4.2.2, υπάρχει ένα κατώφλι για τις τιμές του CRI, το οποίο χρησιμοποιείται για να διακρίνει τα ενδιαφέροντα θέματα από τα μη αντιπροσωπευτικά. Για να καθορίσουμε το όριο για ένα δεδομένο αριθμό θεμάτων, υπολογίζουμε πρώτα τις τιμές των δύο μέτρων (MCP και MWF) για κάθε θέμα. Στη συνέχεια, κανονικοποιούμε τις τιμές για κάθε μέτρο και υπολογίζουμε τις τιμές του CRI.

Οι τιμές που προκύπτουν ταξινομούνται σε αύξουσα σειρά (Σχήμα 5.2) και οι διαφορές των γειτονικών τιμών συγκρίνονται με μια μικρή ποσότητα Δ. Αν η διαφορά είναι μεγαλύτερη από Δ, ορίζουμε το κατώφλι ίσο με το μέσο όρο των δύο τιμών. Τα θέματα πάνω από αυτό το κατώφλι θεωρούνται μη αντιπροσωπευτικά και απορρίπτονται, ενώ θεωρούμε ότι τα θέματα που αντιστοιχούν σε τιμές κάτω από το κατώφλι είναι ενδιαφέροντα.

Η τιμή του Δ εξαρτάται από τη συλλογή των εγγράφων και τον αριθμό των θεμάτων. Για ένα δεδομένο αριθμό θεμάτων, παρατηρούμε άλματα στις τιμές του CRI καθώς πλησιάζουμε τα μη ενδιαφέροντα θέματα, τα οποία μπορούν να



Σχήμα 5.2: Ταξινομημένες τιμές του CRI

βοηθήσουν στον καθορισμό της τιμής του Δ. Για το σύνολο δεδομένων μας, ο αριθμός των τετριμμένων θεμάτων είναι περίπου 20% του συνολικού αριθμού θεμάτων.

5.6 Αυτόματη εξαγωγή επισημάνσεων

Στην ενότητα 4.4.2.3, περιγράφηκε μια μέθοδος αυτόματης δημιουργίας επισημάνσεων. Αυτή η μέθοδος χρησιμοποιεί τις πιο κοινές κατηγορίες άρθρων που αντιστοιχούν στις λέξεις - κλειδιά ενός θέματος ως επισημάνσεις του θέματος. Το αποτέλεσμα αυτής της διαδικασίας εμφανίζεται στον πίνακα 5.3, όπου παρουσιάζεται ο κατάλογος λέξεων-κλειδιών για κάθε θέμα και οι αντίστοιχες επισημάνσεις.

5.7 Εξαγωγή ενδιαφερουσών ομάδων

Τέλος, χρησιμοποιούμε όλες τις μεθόδους που παρουσιάστηκαν προηγουμένως για να βρούμε την τελική ομαδοποίηση και να εξάγουμε τα πιο ενδιαφέροντα

θέματα. Η ομαδοποίηση των χρηστών πραγματοποιήθηκε με τις βέλτιστες τιμές των παραμέτρων ομοιότητας, οι οποίες δίνονται στην ενότητα 5.3.

Στη συνέχεια, υπολογίζουμε το βέλτιστο αριθμό θεμάτων και εκτελούμε το μοντέλο LDA για να ανακτήσουμε τα θέματα στο σύνολο δεδομένων μας. Τα μη αντιπροσωπευτικά θέματα αφαιρούνται (όπως περιγράφηκε στην ενότητα 5.5), και δημιουργούνται οι επισημάνσεις για κάθε μη τετριμμένο θέμα (με τη διαδικασία που παρουσιάστηκε στην ενότητα 4.4.2.3). Τέλος, υπολογίζουμε εκ νέου την τιμή του LTI για όλες τις ομάδες, αυτή τη φορά με τις βέλτιστες παραμέτρους και αφαιρώντας τα ασήμαντα θέματα.

Επομένως, μπορούμε πλέον να ορίσουμε μια νέα μετρική για να ξεχωρίσουμε τις πιο ενδιαφέρουσες ομάδες χρηστών, το *Ενδιαφέρον Τοπικών Κοινοτήτων του Twitter* (Local Twitter Community Interestingness - LTCI), ως εξής:

$$LTCI_r = n_r \cdot LTI_r \tag{5.4}$$

όπου n_r είναι ο αριθμός των χρηστών στην ομάδα r. Υψηλές τιμές του LTCI υποδηλώνουν μεγάλες κοινότητες, που συζητούν ενδιαφέροντα θέματα, τα οποία διαφοροποιούνται από τη γενική συζήτηση.

Ο πίνακας 5.4 απεικονίζει την κανονικοποιημένη τιμή του LTCI για έναν αριθμό αντιπροσωπευτικών ομάδων, τα αντίστοιχα ποσοστά και τις λέξεις - κλειδιά για τα πιο δημοφιλή θέματα σε αυτές τις ομάδες, και τις επισημάνσεις που έχουν εξαχθεί για κάθε θέμα. Τα θέματα που παρουσιάζονται στον πίνακα 5.4 είναι ενδιαφέροντα, με την έννοια ότι η συζήτηση σε αυτές τις ομάδες δεν περιορίζεται στον προγραμματισμό ή την C++, αλλα περιλαμβάνει θεματικές ενότητες όπως videogames ("zelda", "street fighter", "minecraft", "terraria"), ευκαιρίες απασχόλησης σε διαφορετικές χώρες ("Australia", "Mexico", "Romania", "Bulgaria") και talk shows ("Jimmy Fallon", "Jon Stewart").

Πίνακας 5.3: Παραδείγματα θεμάτων και επισημάνσεων

Λέξεις - κλειδιά	Επισημάνσεις
autism love spectrum autistic support people speak accept asperger feel friend part fear story aware xxx depress organ different	Autism, Psychiatric diagnosis
app ipad apple iphone store mac update itunes review dev screen device mini download announce pro touch tablet icon	ITunes, IOS (Apple), Apple Inc. services, Tablet computers
develop source open project web google java android code free api engine service framework javascript github browser website library	Web browsers, Free software, Software licenses, Cross-platform software
git vim github line script commit repo key command branch merge install suck emacs keyboard text shell bash svn	Cross-platform free software, Free software programmed in C, GNU Project software
google facebook twitter social search startup web network tweet share yahoo amazon photo linkedin online engine market launch medium	Internet search engines, Social networking services, Internet companies of the United States, Search engine optimization, American websites, American brands
android google app phone mobile device window samsung develop update nexus galaxy tablet facebook apple product iphone htc buy	Smartphones, Mobile software, Android (operating system) devices, Multinational companies headquartered in the United States
germany german berlin munich frankfurt edinburgh finger europe cold imho cross wouldn birthday london austria flood vienna interpret battery	Capitals in Europe
page email update link account website text site show support screen button click open web image app add follow	Websites
paper data compute research student algorithm interest model analysis learn science universe free process open engine math program library	Scientific method
vettel great race alonso hamilton congratulate silverlight azure vienna webber session service position button store sebastian pole qualify feature	Living people, Formula One World Drivers
linux ubuntu source open kernel desktop debian fedora gnome boot software gnu driver java develop distro android community secure	Free software, Software licenses, Free software programmed in C
game steam play xbox awesome online console epic dead world beta skyrim black stream amaze dragon player valve battlefield	Home video game consoles, Video game terminology

Πίνακας 5.4: Παραδείγματα τιμών του LTCI

LTCI	Ποσοστό θέματος	Λέξεις - κλειδιά	Επισημάνσεις
1	18.65%	party success creative scrapbook join bug cricut adgoggle sell printable pattern shower cash ebat ultimate planner cupcake trial youtube	Arts and crafts, Paper art
	2.06%	play hour super gear fighter war forza halo bundle edit rift street tile zelda dance batman legend arcade adventure	Nintendo Entertainment Analysis and Development games
0.8	50.15%	delphi rad studio develop app builder mobile webinar embarcadero free firemonkey android video register window application device start mac	Windows Mobile, Android (operating system), User interface builders, Mobile operating systems, User interface techniques, ARM operating systems
0.65	4.69%	studio visual code program bit project develop analysis article tool static analyze source programming part bug please error free	User interface builders, Microsoft Visual Studio
	3.94%	sir sydney good maybe cool melbourne cpp australia netlib brisbane night world definite philippin australian suck ftw hey head	Port cities in Australia, Australian capital cities
0.49	18.11%	engine job develop hire senior manager romania bucharest bulgaria sofia software fit anyone good opportunity consult apply automate business	Outlines of countries, Capitals in Europe
	10.47%	mexico operate smb bill guadalajara job city hire unix sql good java anyone analista desarrollador fit slang develop filenet	Cities in Mexico
0.32	36.21%	book ebook develop program save copy review day android learn deal check email send excerpt author app video unleash	Writing occupations, Electronics stubs, Electronic publishing, Dedicated e-book devices, Electronic paper technology
	2.09%	window microsoft app visual net studio develop blog update phone avail build store win msdn preview azure sdk session	.NET Framework
0.26	23.14%	watch late show step fitbit jimmy night toronto stewart jon daily live ferguson mile tonight fallon star general hospital	English-language television programming, Television shows filmed in New York, American late-night television programs

Κεφάλαιο 6

Διανυσματικές Παραστάσεις Κόμβων

Τα τελευταία χρόνια, οι μέθοδοι ενσωμάτωσης (ή ένθεσης) γράφου (graph embedding) έχουν προταθεί ως εναλλακτική στις παραδοσιακές τεχνικές εξόρυξης γράφων. Στόχος είναι η μετατροπή ενός γράφου σε ένα διανυσματικό χώρο χαμηλών διαστάσεων, όπου κάθε κόμβος αντιστοιχεί σε ένα διάνυσμα χαμηλών διαστάσεων. Αυτά τα διανύσματα, που ονομάζονται, επίσης, διανυσματικές παραστάσεις κόμβων (node embeddings) ή διανύσματα χαρακτηριστικών (feature vectors), μπορούν στη συνέχεια να δοθούν ως είσοδοι σε κάποιον αλγόριθμο μηχανικής μάθησης, μετατρέποντας, έτσι, το αρχικό πρόβλημα σε ένα ήδη γνωστό. Επομένως, οι μέθοδοι αυτές είναι χρήσιμες σε μια πληθώρα εφαρμογών του πραγματικού κόσμου, όπως είναι η ταξινόμηση κόμβων, η ανίχνευση κοινοτήτων, η πρόβλεψη συνδέσμου και η οπτικοποίηση δικτύων [110, 18, 40, 35, 108].

6.1 Εισαγωγή

Οι πρόσφατες εργασίες σχετικά με την ενσωμάτωση γράφου χρησιμοποιούν μια παραλλαγή του word2vec για να αντιστοιχίσουν τους κόμβους ενός γράφου σε έναν d-διάστατο διανυσματικό χώρο, διατηρώντας παράλληλα τις σχέσεις μεταξύ των κόμβων. Η ιδέα είναι ότι μετατρέποντας το γράφο σε μια δομή που μοιάζει με προτάσεις, οι κόμβοι αντιμετωπίζονται ως το ισοδύναμο λέξεων και κατά συνέπεια εξάγονται διανυσματικές παραστάσεις κόμβων από το γράφο.

Για τη μετατροπή του γράφου σε μια δομή που μοιάζει με ένα έγγραφο, χρησιμοποιούνται τυχαίοι περίπατοι. Ξεκινώντας από τυχαίους κόμβους και διασχίζοντας το γράφο με τυχαίους περιπάτους, ο γράφος μετατρέπεται σε ένα σύνολο από ακολουθίες κόμβων. Καθώς η μετάβαση γίνεται πάντα από έναν κόμβο σε ένα γειτονικό, οι ακολουθίες αυτές περιέχουν, εμμέσως, πληροφορία

σχετικά με τη δομή του γράφου. Με τη δημιουργία πολλαπλών περιπάτων από κάθε κόμβο του γράφου, μπορεί να δημιουργηθεί ένα αυθαίρετα μεγάλο σύνολο δεδομένων, το οποίο θα ενσωματώνει όλη την πληροφορία για τους κόμβους του γράφου, το οποίο μπορεί στη συνέχεια να χρησιμοποιηθεί για την εξαγωγή των διανυσματικών παραστάσεων. Αυτή η ιδέα διερευνήθηκε στο μοντέλο Deepwalk [73].

Η ιδέα αυτή επεκτάθηκε στο node2vec [38], το οποίο τροποποίησε τη διαδικασία των τυχαίων περιπάτων με στόχο την καλύτερη κωδικοποίηση των διαφόρων ιδιοτήτων του γράφου. Με την προσθήκη δύο υπερπαραμέτρων που εξισορροπούν τη διερεύνηση της κοινοτικής δομής του γράφου με την εξερεύνηση διαφορετικών τμημάτων του δικτύου, οι συγγραφείς κατάφεραν να ξεπεράσουν το Deepwalk στα περισσότερα πειράματα στο πρόβλημα της ταξινόμησης πολλαπλών ετικετών (multi-label classification). Για την επιλογή των τιμών των υπερπαραμέτρων, οι συγγραφείς χρησιμοποίησαν ένα μικρό υποσύνολο κάθε συνόλου δεδομένων (και τις αντίστοιχες ετικέτες) για να διεξάγουν μια αναζήτηση πλέγματος (grid-search) και να βρουν τις βέλτιστες τιμές για το συγκεκριμένο πρόβλημα. Αυτή η διαδικασία, είχε ως αποτέλεσμα το node2vec να διαφοροποιείται από τις υπόλοιπες τεχνικές ενσωμάτωσης (δηλαδή το Deepwalk, αλλά και το word2vec), αφού θεωρείται μια ημι-επιβλεπόμενη τεχνική.

Στο αυτό το κεφάλαιο θα παρουσιαστεί μια νέα μεθοδολογία για την ενσωμάτωση γράφου, βασισμένη σε τυχαίους περιπάτους, η οποία στοχεύει στη χρήση διαφόρων τύπων ομοιότητας μεταξύ των κόμβων για να επιτύχει αντίστοιχα αποτελέσματα με το node2vec, ενώ παραμένει μη επιβλεπόμενη.

6.2 Σχετική έρευνα

Οι πρώτες προσεγγίσεις πάνω στην ενσωμάτωση γράφων μπορούν να θεωρηθούν μέθοδοι μείωσης των διαστάσεων (dimensionality reduction)[50, 6, 87, 102]. Η βασική ιδέα αποτελείται από την κατασκευή μιας αναπαράστασης του γράφου σε μορφή πίνακα (που θα μπορούσε να είναι ο πίνακας γειτνίασης ή ένας πίνακας των ομοιοτήτων/αποστάσεων των κόμβων) και η χρήση μεθόδων γραμμικής άλγεβρας με στόχο την εξαγωγή μιας χαμηλών διαστάσεων αναπαράστασης των κόμβων, όπου οι συνδεδεμένοι κόμβοι παραμένουν κοντά ο ένας στον άλλο στο χώρο ενσωμάτωσης. Οι μέθοδοι αυτές, ωστόσο, είναι απαιτητικές από υπολογιστική άποψη και δεν μπορούν να γενικευθούν καλά σε διαφορετικούς τύπους δικτύων.

Εμπνευσμένες από τις πρόσφατες εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας, και πιο συγκεκριμένα από το μοντέλο Skip-gram [64], πολλές προσεγγίσεις αντιμετωπίζουν ένα γράφο ως «έγγραφο». Μια από αυτές τις προσεγγίσεις, το DeepWalk [73], γενικεύει γλωσσικά μοντέλα για την εκμάθηση λανθανουσών

αναπαραστάσεων των κόμβων ενός γράφου, χρησιμοποιώντας τοπικές πληροφορίες που λαμβάνονται από ομοιόμορφους τυχαίους περιπάτους, οι οποίοι θεωρούνται ισοδύναμοι με προτάσεις.

Ένα άλλο πλαίσιο για την εκμάθηση συνεχών αναπαραστάσεων χαρακτηριστικών για κόμβους σε δίκτυα είναι το node2vec [38]. Σε αντίθεση με τις προηγούμενες προσεγγίσεις, το node2vec χρησιμοποιεί μεροληπτικούς (biased) τυχαίους περιπάτους για την εξερεύνηση του γράφου. Συγκεκριμένα, εισάγει δύο παραμέτρους οι οποίες επηρεάζουν τον περίπατο προς διαφορετικές στρατηγικές εξερεύνησης του δικτύου. Για την εύρεση των βέλτιστων τιμών των παραμέτρων, γίνεται μια αναζήτηση πλέγματος στο σύνολο δεδομένων του εκάστοτε προβλήματος, γεγονός που καθιστά το node2vec μια τεχνική ημι-επιβλεπόμενης μάθησης. Όταν τα δεδομένα δεν είναι επισημασμένα, ο αλγόριθμος χρησιμοποιεί ομοιόμορφους τυχαίους περιπάτους, μετατρέπεται δηλαδή στο Deepwalk.

Μια διαφορετική προσέγγιση που επηρεάζεται από την Επεξεργασία Φυσικής Γλώσσας είναι το NetGlove [39]. Αυτή είναι μια μέθοδος εκμάθησης αναπαραστάσεων των κόμβων που επεκτείνει τον αλγόριθμο Glove [71] από τα έγγραφα στο πλαίσιο των γράφων.

Η σημασία της κωδικοποίησης, όχι μόνο των ακμών μεταξύ των κόμβων, αλλά και των ομοιοτήτων τους, επισημαίνεται σε δύο μεθόδους: τη μέθοδο LINE [99] και τη μέθοδο HOPE [69]. Σε αυτές τις δύο μεθόδους, τα διανύσματα χαρακτηριστικών αναμένεται να διατηρούν διαφορετικές τάξεις της εγγύτητας δικτύου (orders of network proximity). Η εγγύτητα πρώτης τάξης και δεύτερης τάξης των κόμβων ορίζονται ως εξής: δεδομένου ότι το βάρος της ακμής θεωρείται μέτρο ομοιότητας μεταξύ δύο κόμβων, τα βάρη των ακμών ονομάζονται εγγύτητες πρώτης τάξης. Η εγγύτητα δεύτερης τάξης καθορίζεται από την ομοιότητα μεταξύ δύο κόμβων [35].

Η μέθοδος LINE μαθαίνει αναπαραστάσεις χαρακτηριστικών για τους κόμβους του γράφου, ενώ διατηρεί ταυτόχρονα την εγγύτητα πρώτης και δεύτερης τάξης των κόμβων. Ορίζει δύο συναρτήσεις, μία για την εγγύτητα πρώτης τάξης και μία για την εγγύτητα δεύτερης τάξης, και μαθαίνει δύο σύνολα διανυσμάτων αναπαράστασης, ένα για κάθε συνάρτηση. Τέλος, τα δύο διανύσματα που προκύπτουν για κάθε κόμβο από τις δύο μεθόδους συνενώνονται. Παρ΄ ότι η μέθοδος LINE δεν είναι βασισμένη σε τυχαίους περίπατους, συχνά συγκρίνεται και κατατάσσεται μαζί με τα DeepWalk και node2vec [42]. Η μέθοδος HOPE επεκτείνει την LINE προσπαθώντας να διατηρήσει μεγαλύτερης τάξης εγγύτητες, χρησιμοποιώντας γενικευμένη Αποσύνθεση Ιδιαζουσών Τιμών (Singular Value Decomposition - SVD) στον πίνακα ομοιότητας του γράφου.

Τα τελευταία χρόνια, έχουν προταθεί ορισμένοι αλγόριθμοι οι οποίοι προσφέρουν μια άλλη οπτική στο πρόβλημα της ενσωμάτωσης γράφου. Πιο συγκεκριμένα, οι συγκεκριμένες προσεγγίσεις έχουν ως στόχο την κωδικοποίηση των σχέσεων ισοδυναμίας ρόλου ανάμεσα στους κόμβους. Δύο κόμβοι έχουν

τον ίδιο ρόλο μέσα σε ένα δίκτυο όταν οι γειτονιές τους είναι δομικά όμοιες. Παρόλο που η μελέτη των δομικών ιδιοτήτων των κόμβων ξεφεύγει από τη σκοπιά της παρούσας εργασίας, αξίζει να γίνει μια αναφορά σε τέτοιου είδους προσεγγίσεις για λόγους πληρότητας. Μια τέτοια ενδιαφέρουσα προσέγγιση, η οποία είναι βασισμένη σε τυχαίους περιπάτους, είναι το struc2vec [29]. Αυτός ο αλγόριθμος μαθαίνει λανθάνουσες αναπαραστάσεις για τη δομική ταυτότητα των κόμβων, κωδικοποιώντας τις δομικές ομοιότητες των κόμβων. Μια πιο εξελιγμένη εκδοχή αυτού του αλγορίθμου είναι το struc2vec++ [97], το οποίο γενικεύει το struc2vec σε κατευθυνόμενους και σταθμισμένους γράφους.

Μια ενδιαφέρουσα προσέγγιση είναι η μέθοδος HARP (Hierarchical Representation Learning for Networks) [19]. Η συγκεκριμένη μέθοδος κάνει συμπίεση (compress) στο γράφο εισόδου πριν την εφαρμογή μιας τεχνικής ενσωμάτωσης, συγχωνεύοντας κάποιους κόμβους, άρα μειώνοντας το μέγεθος του γράφου. Πρακτικά, αποτελεί μια τεχνική προ-επεξεργασίας του γράφου, η οποία διευκολύνει τη διαχείριση μεγάλων δικτύων, ενώ ταυτόχρονα καταφέρνει να βελτιώσει την απόδοση των διαφόρων τεχνικών ενσωμάτωσης.

Η παρούσα εργασία επικεντρώνεται στις προσεγγίσεις που βασίζονται σε τυχαίους περίπατους. Παρ΄ όλα αυτά, υπάρχει πληθώρα αλγορίθμων που προσεγγίζουν το πρόβλημα με διαφορετικό τρόπο [42, 41].

6.3 Περιγραφή της μεθοδολογίας

Οι τυχαίοι περίπατοι αποτελούν ένα αποτελεσματικό εργαλείο στην εξερεύνηση των δικτύων. Είναι ιδιαίτερα αποτελεσματικοί όταν ο γράφος είναι πολύ μεγάλος ή όταν υπάρχει πληροφορία μόνο για ορισμένα τμήματα του γράφου, καθώς οι τυχαίοι περίπατοι χρησιμοποιούν μόνο τοπικές πληροφορίες του δικτύου.

Πολλοί δημοφιλείς αλγόριθμοι ανίχνευσης κοινοτήτων βασίζονται σε τυχαίους περίπατους [31, 18, 110]. Ο λόγος είναι ότι όταν ένας γράφος έχει ισχυρή κοινοτική δομή, οι κόμβοι στο εσωτερικό των κοινοτήτων είναι πιο πυκνά συνδεδεμένοι μεταξύ τους παρά με το υπόλοιπο δίκτυο, αναγκάζοντας τους τυχαίους περιπατητές να περάσουν περισσότερο χρόνο μέσα στις κοινότητες.

Σε αυτή την ενότητα παρουσιάζεται μια νέα, μη επιβλεπόμενη προσέγγιση στην ενσωμάτωση γράφων, βασισμένη σε αυτή τη γνωστή ιδιότητα των τυχαίων περιπάτων. Σε αντίθεση με τις προηγούμενες προσεγγίσεις, οι οποίες λαμβάνουν υπόψη μόνο τις ακμές ενός γράφου κατά την εξερεύνηση του γράφου μέσω τυχαίων περιπάτων, η προτεινόμενη μεθοδολογία λαμβάνει επίσης υπόψη τις ομοιότητες μεταξύ των κόμβων.

Η διαίσθηση πίσω από τα παραπάνω βασίζεται σε μια απλή ιδέα: σε κάθε βήμα του τυχαίου περίπατου, αντί να επιλέγουμε τυχαία μεταξύ των γειτόνων

ενός κόμβου, να αναγκάσουμε τον τυχαίο περίπατο να μετακινείται σε κόμβους που είναι παρόμοιοι με τον τρέχοντα κόμβο. Αυτό έχει ως αποτέλεσμα αναπαραστάσεις σε ένα διανυσματικό χώρο, που δεν περιλαμβάνουν μόνο πληροφορία σχετική με τις ακμές μεταξύ των κόμβων, αλλά κωδικοποιούν και τις ομοιότητες μεταξύ κάθε ζεύγους κόμβων ως προς κάποια ιδιότητα.

Μέσω της μεταβολής της πιθανότητας επίσκεψης κάθε κόμβου, και θέτοντας την ανάλογη με την ομοιότητα μεταξύ των δύο κόμβων, ελαχιστοποιείται η πιθανότητα διάσχισης ακμών που βρίσκονται ανάμεσα σε κοινότητες. Ως αποτέλεσμα, η προτεινόμενη μεθοδολογία αναμένεται να έχει καλύτερη απόδοση σε γράφους με ισχυρή κοινοτική δομή.

6.3.1 Ορισμός του προβλήματος

Έστω $G = (V, E)$ ένας γράφος, όπου το V αντιστοιχεί στους κόμβους και το E στις ακμές, και $E \subseteq (V \times V)$. Ο στόχος είναι ο προσδιορισμός μιας συνάρτησης $f : u \to \boldsymbol{v} \in \mathbb{R}^d$, $\forall u \in V$, όπου $d \ll |V|$. Επομένως, η συνάρτηση f αντιστοιχίζει κάθε κόμβο του γράφου σε ένα διάνυσμα χαρακτηριστικών $\boldsymbol{v}$ χαμηλών διαστάσεων. Η ιδέα είναι να ελαχιστοποιηθεί η απόσταση μεταξύ των κόμβων στο διανυσματικό χώρο, εάν οι κόμβοι είναι συνδεδεμένοι ή παρόμοιοι μεταξύ τους. Ως αποτέλεσμα, η f αναμένεται να διατηρεί τις συνδέσεις μεταξύ των κόμβων του G, καθώς και την πληροφορία σχετικά με την ομοιότητα των κόμβων.

Σε αυτή την εργασία, οι γράφοι θεωρούνται ότι είναι μη σταθμισμένοι (δεν έχουν βάρη στις ακμές). Ωστόσο, αυτή η μεθοδολογία μπορεί εύκολα να εφαρμοστεί σε σταθμισμένους γράφους με μικρές προσαρμογές.

6.3.2 Περιγραφή του συστήματος

Οι προσεγγίσεις που βασίζονται σε τυχαίους περιπάτους συνήθως αποτελούνται από δύο κύρια στοιχεία, μια γεννήτρια τυχαίων περιπάτων και μια διαδικασία ενημέρωσης [73, 38]. Η προτεινόμενη μεθοδολογία ακολουθεί την ίδια αρχή, με την εξαίρεση ότι εξετάζονται τρεις διαφορετικές γεννήτριες τυχαίων περιπάτων. Κάθε γεννήτρια εφαρμόζει μια διαφορετική διαδικασία εξερεύνησης του δικτύου, διατηρώντας διαφορετικές ιδιότητες. Επιπλέον, προτείνεται ένας τρόπος συνδυασμού των αποτελεσμάτων αυτών των τριών προσεγγίσεων.

6.3.2.1 Γεννήτριες τυχαίων περιπάτων

Η μεθοδολογία επικεντρώνεται σε τρεις διαφορετικούς τρόπους εξερεύνησης του δικτύου, που έχουν ως αποτέλεσμα τρεις διαφορετικές γεννήτριες τυχαίων περιπάτων. Κάθε μία από αυτές διατηρεί διαφορετικές ιδιότητες του δικτύου.

Οι γεννήτριες λαμβάνουν έναν γράφο G ως είσοδο και επιστρέφουν ένα σύνολο τυχαίων περιπάτων. Υπάρχουν δύο υπερπαράμετροι που εμπλέκονται σε αυτή τη διαδικασία, ο αριθμός των περιπάτων ανά κόμβο t και ο αριθμός των βημάτων ανά περίπατο s.

6.3.2.1.1 Ομοιόμορφοι (uniform) τυχαίοι περίπατοι (U): Η πρώτη γεννήτρια χρησιμοποιεί ομοιόμορφους τυχαίους περίπατους για την εξερεύνηση του γράφου. Σε κάθε βήμα, η γεννήτρια επισκέπτεται τον επόμενο κόμβο επιλέγοντας ομοιόμορφα από τους γείτονες του τρέχοντος κόμβου. Η πιθανότητα μετακίνησης από τον κόμβο u_i στον u_j σε ένα βήμα δίνεται από την ακόλουθη εξίσωση:

$$Pr(u_j|u_i) = \begin{cases} \frac{1}{Z_i}, & \text{αν } (u_i, u_j) \in E \\ 0, & \text{αλλιώς} \end{cases} \tag{6.1}$$

όπου Z_i είναι μια σταθερά κανονικοποίησης, ώστε να ισχύει ότι:

$$\sum_{u_j \in E} Pr(u_j|u_i) = 1 \tag{6.2}$$

Βασικά, αυτή η γεννήτρια είναι ίδια με αυτή του μοντέλου Deepwalk. Η γεννήτρια αυτή αποσκοπεί στη διατήρηση της πληροφορίας σχετικά με τις συνδέσεις (τις ακμές) μεταξύ των κόμβων.

6.3.2.1.2 Μη ομοιόμορφοι (non-uniform) τυχαίοι περίπατοι σε όμοιους γειτονικούς κόμβους (S_{nbr}): Η δεύτερη γεννήτρια επιλέγει τον επόμενο κόμβο με βάση την ομοιότητα μεταξύ του τρέχοντος κόμβου και κάθε γείτονα του. Ένας πιο επίσημος ορισμός της ομοιότητας μεταξύ δύο κόμβων θα δοθεί στην ενότητα 6.3.2.2. Σε κάθε βήμα, η πιθανότητα μετακίνησης σε γειτονικό κόμβο δίνεται από την ακόλουθη εξίσωση:

$$Pr(u_j|u_i) = \begin{cases} \frac{sim(u_i,u_j)}{Z_i}, & \text{αν } (u_i, u_j) \in E \\ 0, & \text{αλλιώς} \end{cases} \tag{6.3}$$

όπου $sim(u_i, u_j)$ είναι μία μετρική που υπολογίζει την ομοιότητα μεταξύ των κόμβων u_i και u_j.

6.3.2.1.3 Μη ομοιόμορφοι (non-uniform) τυχαίοι περίπατοι σε οποιονδήποτε όμοιο κόμβο (S_{any}): Η τρίτη γεννήτρια επίσης χρησιμοποιεί μη ομοιόμορφους τυχαίους περίπατους, ωστόσο η επιλογή του επόμενου

κόμβου σε κάθε βήμα δεν περιορίζεται στο σύνολο γειτόνων του τρέχοντος κόμβου. Ο τυχαίος περίπατος μπορεί να μετακινηθεί σε οποιονδήποτε από τους κόμβους, με βάση την ομοιότητα μεταξύ των δύο κόμβων:

$$Pr(u_j|u_i) = \frac{sim(u_i, u_j)}{Z_i}, \quad i \neq j \tag{6.4}$$

Αυτή η γεννήτρια έχει ως σκοπό να διατηρήσει τις ομοιότητες μεταξύ των κόμβων.

Η διαίσθηση πίσω από τη δεύτερη και την τρίτη γεννήτρια είναι ότι, αναγκάζοντας τον τυχαίο περίπατο να μετακινηθεί σε κόμβους παρόμοιους με τον τρέχοντα κόμβο, ελαχιστοποιούνται οι πιθανότητες διάσχισης ακμών που βρίσκονται μεταξύ των κοινοτήτων, επομένως κάθε περίπατος περιορίζεται στα όρια μιας κοινότητας.

Ειδικά, για την τρίτη γεννήτρια, ο περιορισμός ότι ένας τυχαίος περίπατος μπορεί να μετακινηθεί μόνο σε γειτονικούς κόμβους αφαιρείται. Η ιδέα προέρχεται από πολλές παραδοσιακές μεθόδους συσταδοποίησης, οι οποίες υπολογίζουν την ομοιότητα μεταξύ κάθε ζεύγους κορυφών σε σχέση με κάποια ιδιότητα, χωρίς να εξετάζεται εάν συνδέονται με ακμή ή όχι.

6.3.2.2 Μετρικές ομοιότητας

Υπάρχουν πολλές μετρικές ομοιότητας που χρησιμοποιούν πληροφορία σχετική με τη γειτονιά των κόμβων για να υπολογίσουν την ομοιότητα μεταξύ τους. Σε αυτή την ενότητα θα περιγραφούν πέντε τέτοιες μετρικές, οι οποίες στη συνέχεια θα χρησιμοποιηθούν για την παραγωγή τυχαίων περιπάτων. Κάθε μία από τις παρακάτω μετρικές μπορεί να υποκαταστήσει την $sim(u_i, u_j)$ στις εξισώσεις 6.3 και 6.4.

6.3.2.2.1 Αριθμός κοινών γειτόνων Ένας τρόπος για τον προσδιορισμό της ομοιότητας μεταξύ των κόμβων είναι ο υπολογισμός του αριθμού των κοινών γειτόνων τους:

$$cn(u_i, u_j) = |\Gamma(u_i) \cap \Gamma(u_j)| \tag{6.5}$$

όπου $\Gamma(u_i)$ και $\Gamma(u_j)$ είναι οι γειτονιές των κόμβων u_i και u_j, αντίστοιχα, επομένως $|\Gamma(u_i) \cap \Gamma(u_j)|$ είναι ο αριθμός των κοινών γειτόνων των u_i και u_j.

6.3.2.2.2 Δείκτης Jaccard Μια άλλη γνωστή μετρική ομοιότητας είναι ο δείκτης Jaccard, ο οποίος δίνεται στην ακόλουθη εξίσωση:

$$J(u_i, u_j) = \frac{|\Gamma(u_i) \cap \Gamma(u_j)|}{|\Gamma(u_i) \cup \Gamma(u_j)|} \tag{6.6}$$

όπου $|\Gamma(u_i) \cap \Gamma(u_j)|$ είναι ο αριθμός κοινών γειτόνων των u_i και u_j και $|\Gamma(u_i) \cup \Gamma(u_j)|$ είναι ο συνολικός αριθμός των γειτόνων των u_i και u_j. Αυτή η μετρική είναι παρόμοια με την προηγούμενη, με τη διαφορά ότι είναι κανονικοποιημένη ως προς το συνολικό αριθμό γειτόνων των u_i και u_j, γεγονός που σημαίνει ότι πλέον δεν ευνοεί τους κόμβους με μεγάλο αριθμό γειτόνων.

6.3.2.2.3 Ευκλείδεια απόσταση Η ομοιότητα μπορεί επίσης να συναχθεί από τον πίνακα γειτνίασης μεταξύ κορυφών. Ένα πιθανό μέτρο απόστασης μπορεί να οριστεί ως:

$$d(u_i, u_j) = \sqrt{\sum_{k=1}^{n} (A_{ik} - A_{jk})^2} \tag{6.7}$$

όπου $\mathbf{A}$ είναι ο πίνακας γειτνίασης του γράφου και n είναι ο αριθμός κόμβων. Αυτό το μέτρο μπορεί να υπολογιστεί είτε με τις γραμμές είτε με τις στήλες του πίνακα γειτνίασης. Δεδομένου ότι αυτό είναι στην πραγματικότητα ένα μέτρο ανομοιότητας και παρόμοιοι κόμβοι αναμένεται να είναι κοντά ο ένας στον άλλο, μεγαλύτερες τιμές αυτού του μέτρου αντιστοιχούν σε κόμβους που διαφέρουν περισσότερο.

6.3.2.2.4 Ομοιότητα συνημιτόνου Ένα άλλο δημοφιλές χωρικό μέτρο είναι η ομοιότητα συνημιτόνου, η οποία μπορεί να υπολογιστεί με τον ακόλουθο τρόπο:

$$\rho(u_i, u_j) = \frac{\sum_{k=1}^{n} A_{ik} A_{jk}}{\sqrt{\sum_{k=1}^{n} A_{ik}^2} \sqrt{\sum_{k=1}^{n} A_{jk}^2}} \tag{6.8}$$

6.3.2.2.5 Συσχέτιση Pearson Ένα άλλο μέτρο είναι η συσχέτιση Pearson, η οποία μπορεί να υπολογιστεί μεταξύ των γραμμών ή των στηλών του πίνακα γειτνίασης:

$$C(u_i, u_j) = \frac{\sum_{k=1}^{n} (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n \sigma_i \sigma_j} \tag{6.9}$$

όπου μ_i είναι οι μέσες τιμές και σ_i είναι οι τυπικές αποκλίσεις, που δίνονται από τις ακόλουθες εξισώσεις:

$$\mu_i = \frac{\sum_k A_{ik}}{n} \tag{6.10}$$

$$\sigma_i = \sqrt{\frac{\sum_k (A_{ik} - \mu_i)^2}{n}} \tag{6.11}$$

Είναι σημαντικό να τονιστεί ότι η λίστα των μετρικών ομοιότητας που παρουσιάζονται εδώ δεν είναι πλήρης. Ενώ υπάρχουν πολλοί διαφορετικοί τρόποι για να υπολογιστεί η ομοιότητα μεταξύ δύο κόμβων (π.χ. ο αριθμός των ανεξάρτητων διαδρομών ως προς τους κόμβους ή τις ακμές μεταξύ δύο κόμβων, η κεντρικότητα ενδιαμεσότητας ακμών (edge betweenness centrality) για ένα ζευγάρι γειτονικών κόμβων), οι μετρικές επιλέχθηκαν με βάση δύο κριτήρια. Οι μετρικές θα πρέπει να χρησιμοποιούν μόνο τοπικές πληροφορίες του γράφου, ενώ ο υπολογισμός των ομοιοτήτων θα πρέπει να είναι αποδοτικός ως προς το χρόνο.

6.3.2.3 Διαδικασία ενημέρωσης

Για την παραγωγή των διανυσμάτων αναπαράστασης από τους τυχαίους περίπατους, χρησιμοποιείται ο αλγόριθμος Skip-gram. Όπως εξηγήθηκε νωρίτερα, αποτελείται από ένα νευρωνικό δίκτυο με ένα κρυφό επίπεδο, το οποίο παράγει ένα διανυσματικό χώρο για λέξεις όταν δίνεται ένα μεγάλο σώμα κειμένων ως είσοδος. Κάθε λέξη αντιστοιχίζεται σε ένα διάνυσμα στο χώρο, ενώ οι λέξεις που εμφανίζονται στο ίδιο πλαίσιο στα κείμενα, βρίσκονται κοντά στο διανυσματικό χώρο.

Στην περίπτωση των γράφων, οι τυχαίοι περίπατοι αντιμετωπίζονται ως ισοδύναμοι των προτάσεων και ένα μετακινούμενο παράθυρο επιλέγει τους κόμβους που εμφανίζονται στο ίδιο πλαίσιο. Οι δύο παράμετροι που εμπλέκονται στη διαδικασία ενημέρωσης είναι το μέγεθος του παραθύρου w και οι διαστάσεις d των διανυσμάτων που προκύπτουν. Το μοντέλο Skip-gram μπορεί να εκπαιδευτεί είτε με ιεραρχική softmax είτε με αρνητική δειγματοληψία, με την τελευταία να είναι αυτή που χρησιμοποιείται σε αυτή τη μελέτη, καθώς οδηγεί σε ταχύτερη εκπαίδευση.

6.3.2.4 Συνδυασμός των διαφόρων στρατηγικών

Όπως εξηγήθηκε προηγουμένως, αυτή η μεθοδολογία χρησιμοποιεί τρεις διαφορετικές γεννήτριες τυχαίων περιπάτων, κάθε μία εκ των οποίων διατηρεί διαφορετικές ιδιότητες του δικτύου. Ο συνδυασμός κάθε γεννήτριας με τη διαδικασία ενημέρωσης έχει ως αποτέλεσμα διαφορετικό σύνολο χαρακτηριστικών. Ως τελευταίο βήμα στη μεθοδολογία, τα αποτελέσματα από τις διαφορετικές διαδικασίες εξερεύνησης μπορούν να συνδυαστούν. Αυτό μπορεί να συμβεί με τη

συνένωση (concatenation), για κάθε κόμβο, των διανυσμάτων ενσωμάτωσης που προέκυψαν από τη χρήση διαφορετικών γεννητριών τυχαίων περιπάτων.

Είναι σημαντικό να σημειωθεί ότι σε όλα τα πειράματα που θα παρουσιαστούν στο επόμενο κεφάλαιο, τα διανύσματα χαρακτηριστικών έχουν το ίδιο μέγεθος, παρά τη συνένωση. Αυτό επιτυγχάνεται ρυθμίζοντας το μέγεθος από κάθε διάνυσμα που συνενώνεται, το οποίο είναι ίσο με $d/2$ για κάθε διάνυσμα, όταν συνδυάζονται δύο διαδικασίες εξερεύνησης και $d/3$ όταν συνδυάζονται και οι τρεις διαδικασίες εξερεύνησης.

Κεφάλαιο 7

Διανυσματικές Παραστάσεις Κόμβων - Πειραματικό μέρος

7.1 Σύνολα δεδομένων

Ως πρώτο βήμα, η προτεινόμενη μεθοδολογία αξιολογείται σε τεχνητά δίκτυα που παράγονται από τον αλγόριθμο Lancichinetti-Fortunato-Radicchi (LFR) [56]. Ο αλγόριθμος LFR παράγει τεχνητά δίκτυα που έχουν a priori γνωστές κοινότητες, επομένως μπορούν να χρησιμοποιηθούν για τη σύγκριση διαφορετικών μεθόδων ανίχνευσης κοινοτήτων.

Ο αλγόριθμος έχει μια παράμετρο μίξης (mixing parameter) μ, η οποία αντιπροσωπεύει την ποσότητα του θορύβου στο δίκτυο. Κάθε κόμβος μοιράζεται ένα κλάσμα $1-\mu$ των συνδέσεών του με κόμβους ίδιας κοινότητας και ένα κλάσμα μ με άλλους κόμβους του δικτύου. Όταν $\mu = 0$ όλες οι συνδέσεις είναι ανάμεσα σε κόμβους που ανήκουν στην ίδια κοινότητα, ενώ όταν $\mu = 1$ είναι μεταξύ κόμβων που ανήκουν σε διαφορετικές κοινότητες.

Εννέα οικογένειες δικτύων δημιουργούνται με αυτό τον τρόπο, με διαφορετικές τιμές της παραμέτρου μίξης μ, που κυμαίνονται από 0.1 έως 0.9, με βήμα 0.1. Κάθε οικογένεια δικτύων αποτελείται από 10 δίκτυα. Μετά την εξαγωγή των αποτελεσμάτων, υπολογίζεται ο μέσος όρος για τα δίκτυα της ίδιας οικογένειας, ώστε να προστεθεί περισσότερη ευρωστία στην πειραματική διαδικασία. Τα τεχνητά δίκτυα έχουν 1000 κόμβους το καθένα, με μέσο βαθμό ίσο με 20 και μέγιστο βαθμό 50.

Το δεύτερο μέρος της αξιολόγησης περιλαμβάνει σύνολα δεδομένων του πραγματικού κόσμου. Συγκεκριμένα, η προτεινόμενη μεθοδολογία αξιολογείται με βάση τα παρακάτω σύνολα δεδομένων:

- Το BlogCatalog [100] είναι ένα δίκτυο κοινωνικών σχέσεων μεταξύ blog-

gers, το οποίο έχει εξαχθεί από την αντίστοιχη ιστοσελίδα[1]. Περιλαμβάνει τόσο το δίκτυο φιλίας όσο και την πληροφορία για τις ομάδες στις οποίες είναι μέλη οι κόμβοι.

- Το Protein-Protein Interactions (PPI) [15] είναι ένα δίκτυο βιολογικών αλληλεπιδράσεων μεταξύ πρωτεϊνών στους ανθρώπους.

- Το Wikipedia [62] είναι ένα δίκτυο συν-εμφάνισης λέξεων, οι οποίες εμφανίζονται στο πρώτο εκατομμύριο bytes του Wikipedia dump.

Ο πίνακας 7.1 συνοψίζει τα χαρακτηριστικά των συνόλων δεδομένων. Κάθε ένα από τα δίκτυα επιλέχθηκαν στο [38] επειδή παρουσιάζουν διαφορετικά μείγματα ομοφιλίας και ισοδυναμίας ρόλων.

Όσον αφορά το σύνολο δεδομένων BlogCatalog, το οποίο αποτελείται από τις σχέσεις φιλίας μεταξύ bloggers και στο οποίο οι επισημάνσεις για τους κόμβους αντιπροσωπεύουν τα ενδιαφέροντα κάθε χρήστη, αναμένεται ότι οι κόμβοι θα παρουσιάζουν ισχυρές σχέσεις βασισμένες στην ομοφιλία, επομένως οι επισημάνσεις θα αποκαλύπτουν την κοινοτική δομή του δικτύου.

Το σύνολο δεδομένων Protein-Protein Interactions (PPI) διαθέτει και τα δύο είδη ισοδυναμιών. Οι πρωτεΐνες μπορούν να εκτελούν είτε παρόμοιες, είτε συμπληρωματικές λειτουργίες με τις γειτονικές τους πρωτεΐνες, δημιουργώντας έτσι σχέσεις μεταξύ των κόμβων που βασίζονται στην ομοφιλία ή στην ισοδυναμία ρόλων.

Τέλος, το σύνολο δεδομένων Wikipedia είναι ένα πυκνό δίκτυο λέξεων, με τις ετικέτες να είναι το μέρος του λόγου κάθε λέξης. Δεδομένου ότι το δίκτυο είναι πυκνό, οι λέξεις με την ίδια ετικέτα θα συνδέονται μεταξύ τους. Ταυτόχρονα, κυρίως επειδή μέσα σε μια πρόταση εμφανίζονται συντακτικά και γραμματικά μοτίβα, οι ισοδυναμίες ρόλων θα εμφανίζονται επίσης μέσα στο δίκτυο.

Πίνακας 7.1: Επισκόπηση των συνόλων δεδομένων

Όνομα	$\|V\|$	$\|E\|$	Αριθμός επισημάνσεων
BlogCatalog	10,312	333,983	39
Protein-Protein Interactions	3,890	38,739	50
Wikipedia	4,777	184,812	40

[1]http://www.blogcatalog.com

7.2 Πειραματική διαδικασία

Η προτεινόμενη μεθοδολογία αξιολογείται στο πρόβλημα της ταξινόμησης πολλαπλών ετικετών. Για να είναι εύκολη η σύγκριση με τις άλλες μεθόδους, χρησιμοποιείται ακριβώς η ίδια πειραματική διαδικασία όπως στο [73]. Τα δεδομένα χωρίζονται σε σύνολο εκπαίδευσης (training set) και ελέγχου (test set) με τυχαία δειγματοληψία ενός τμήματος των κόμβων, οι οποίοι θα χρησιμοποιηθούν ως σύνολο εκπαίδευσης, ενώ οι υπόλοιποι χρησιμοποιούνται ως σύνολο ελέγχου. Η διαδικασία αυτή επαναλαμβάνεται 10 φορές. Στη συνέχεια, τα δεδομένα εκπαίδευσης δίνονται ως είσοδος σε έναν ένα-έναντι-υπολοίπων (one-vs-rest) ταξινομητή λογιστικής παλινδρόμησης (logistic regression), ο οποίος έχει επεκταθεί για να επιστρέφει τις πιο πιθανές επισημάνσεις. Τα αποτελέσματα αξιολογούνται με τη χρήση των μετρικών Macro-F1 και Micro-F1.

Η απόδοση των προτεινόμενων μοντέλων αξιολογείται σε σύγκριση με το Deepwalk και το node2vec. Όπως αναφέρθηκε και προηγουμένως, η συγκεκριμένη εργασία επικεντρώνεται και συγκρίνεται με προσεγγίσεις που βασίζονται σε τυχαίους περίπατους. Ο λόγος είναι ότι η απόδοση των συγκεκριμένων μοντέλων επηρεάζεται σε μεγάλο βαθμό από τον αριθμό των δειγμάτων που δίνονται ως είσοδος στο μοντέλο Skip-gram (ο οποίος επηρεάζεται από την επιλογή των παραμέτρων, δηλαδή τον αριθμό περιπάτων ανά κόμβο, τον αριθμό βημάτων ανά περίπατο, το μέγεθος παραθύρου και τον αριθμό εποχών), επομένως η σύγκριση με παρόμοιες μεθόδους είναι ευκολότερη.

Μια μέθοδος η οποία δεν είναι βασισμένη σε τυχαίους περίπατους, αλλά συχνά κατατάσσεται μαζί με τα DeepWalk και node2vec [42] είναι η μέθοδος LINE. Παρ΄ όλα αυτά, τα αποτελέσματα για τη συγκεκριμένη μέθοδο δεν συμπεριλήφθηκαν στην παρούσα εργασία. Ο λόγος είναι ότι σύμφωνα με τη δημοσίευση του node2vec, όταν παράγεται ίδιος αριθμός δειγμάτων κατά την εκτέλεση, τα δύο μοντέλα (Deepwalk και node2vec) δίνουν καλύτερα αποτελέσματα από τη μέθοδο LINE [38].

Προκειμένου να εξασφαλίσουμε ότι η σύγκριση είναι δίκαιη, παράγεται ίσος αριθμός δειγμάτων εκπαίδευσης για όλες τις μεθόδους, ενώ η βελτιστοποίηση εκτελείται για μία μόνο εποχή για όλα τα μοντέλα. Οι παράμετροι που χρησιμοποιούνται για το Deepwalk, node2vec και τα προτεινόμενα μοντέλα (όταν χρησιμοποιείται μόνο μία διαδικασία εξερεύνησης) είναι:

- διαστάσεις διανύσματος:
  - $d = 12$ για τα τεχνητά δίκτυα
  - $d = 300$ για τα δίκτυα του πραγματικού κόσμου
- περίπατοι ανα κόμβο: $t = 10$

- βήματα ανά περίπατο: $s = 80$
- παράθυρο: $w = 10$

Όταν συνδυάζονται δύο διαδικασίες εξερεύνησης, οι διαστάσεις του διανύσματος ορίζονται σε $d = 6$ για τα τεχνητά δίκτυα και σε $d = 150$ για τα σύνολα δεδομένων πραγματικού κόσμου, για κάθε διαδικασία, ενώ όταν συνδυάζονται τρεις διαδικασίες, οι διαστάσεις είναι $d = 4$ και $d = 100$, αντίστοιχα.

Τέλος, δεδομένου ότι το node2vec έχει δύο πρόσθετες παραμέτρους, πρέπει να σημειωθεί ότι σε όλα τα πειράματα χρησιμοποιήθηκαν οι βέλτιστες τιμές για αυτές τις παραμέτρους, οι οποίες υπολογίστηκαν με αναζήτηση πλέγματος, ακριβώς όπως περιγράφεται στο [38].

7.3 Πειραματικά αποτελέσματα

7.3.1 Τεχνητά δίκτυα

Ως πρώτο βήμα για την αξιολόγηση των διαφορετικών γεννητριών τυχαίων περιπάτων, η μεθοδολογία δοκιμάζεται στους γράφους LFR. Τα σχήματα 7.1 και 7.2 παρουσιάζουν τα αποτελέσματα όταν χρησιμοποιούνται οι διαδικασίες εξερεύνησης S_{nbr} και S_{any}, αντίστοιχα. Τα αποτελέσματα της πρώτης γεννήτριας (U) παραλείπονται, καθώς συμπίπτουν με τα αποτελέσματα του Deepwalk, τα οποία συμπεριλαμβάνονται και στα δύο σχήματα. Η απόδοση του συστήματος απεικονίζεται για διαφορετικές τιμές της παραμέτρου μίξης μ, όπου το ποσοστό των δεδομένων εκπαίδευσης ισούται με 10%. Τα αποτελέσματα της χρήσης των διαφορετικών μετρικών ομοιότητας συγκρίνονται με τις δύο μεθοδολογίες.

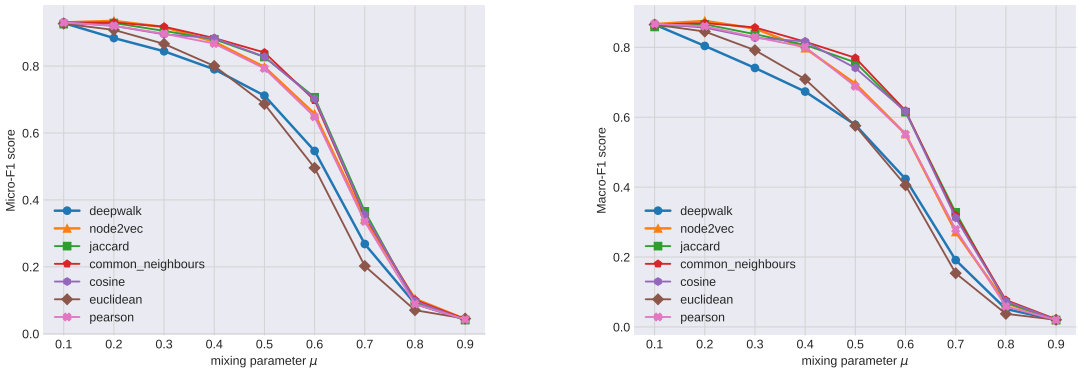
Και στις δύο περιπτώσεις, είναι προφανές ότι καθώς η τιμή της παραμέτρου μίξης μ αυξάνεται, η απόδοση όλων των συστημάτων μειώνεται. Όπως εξηγήθηκε προηγουμένως, όταν η τιμή του μ είναι κοντά στο 1, οι περισσότερες ακμές του γράφου συνδέουν κόμβους που ανήκουν σε διαφορετικές κοινότητες. Κατά συνέπεια, αναμένεται αυτή η μείωση της απόδοσης.

Στην περίπτωση της S_{nbr}, τα αποτελέσματα για όλες σχεδόν τις μετρικές ομοιότητας είναι αντίστοιχα ή καλύτερα σε σχέση εκείνα του node2vec, ενώ υπερβαίνουν του DeepWalk, ακόμη και όταν οι τιμές της παραμέτρου μίξης μ είναι κοντά στο 1. Η μόνη εξαίρεση είναι όταν χρησιμοποιείται η ευκλείδεια απόσταση, όπου η απόδοση είναι συγκρίσιμη με του DeepWalk.

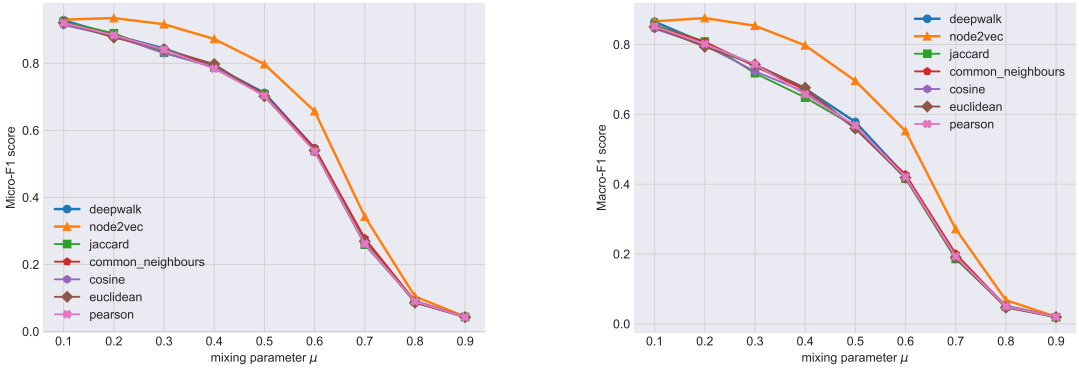
Αντίθετα, η διαδικασία εξερεύνησης S_{any} δεν παρουσιάζει καλά αποτελέσματα, αλλά έχει παρόμοια αποτελέσματα με τον αλγόριθμο DeepWalk. Δεδομένου ότι τα τεχνητά δίκτυα LFR είναι ιδανικά δίκτυα, όπου οι επισημάνσεις των κόμβων αντιπροσωπεύουν τις κοινότητές τους, πράγμα που σημαίνει ότι οι κόμβοι με την ίδια ετικέτα συνδέονται σε κάποιο βαθμό μεταξύ τους και ότι η

τρίτη γεννήτρια δεν λαμβάνει υπόψη τις πληροφορίες που παρέχονται από τις συνδέσεις μεταξύ των κόμβων, αυτά τα χειρότερα αποτελέσματα μπορούν να δικαιολογηθούν.

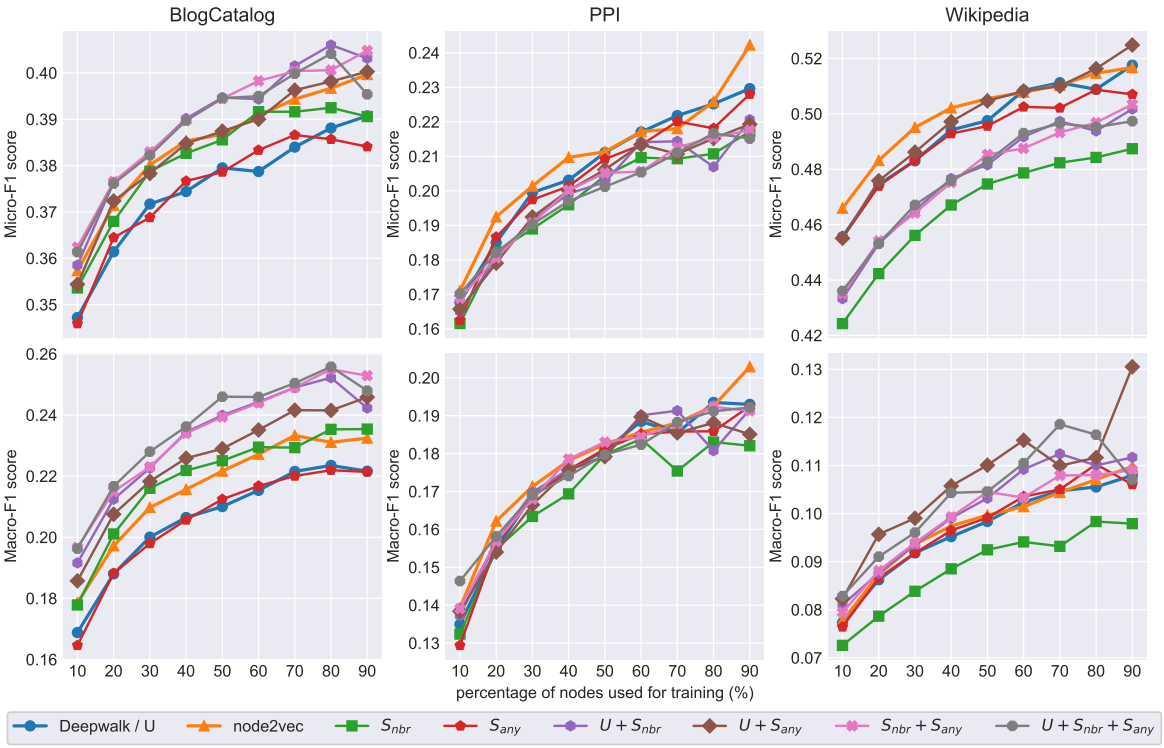
Θα πρέπει να σημειωθεί ότι αυτά τα δίκτυα είναι «ιδανικά» και δεν μπορούν να εκφέρουν επαρκώς την πολυπλοκότητα των πραγματικών δικτύων. Ωστόσο, όπως υποδεικνύεται από την απόδοση της S_{nbr}, η προσθήκη πληροφορίας σχετικά με την ομοιότητα των κόμβων μπορεί να οδηγήσει σε καλύτερες απεικονίσεις. Αυτό θα εξεταστεί στην επόμενη ενότητα, όπου η απόδοση της προτεινόμενης μεθοδολογίας θα αξιολογηθεί σε μη ιδανικές συνθήκες.



Σχήμα 7.1: Αποτελέσματα της δεύτερης γεννήτριας τυχαίων περιπάτων (S_{nbr}), για διαφορετικές μετρικές ομοιότητας, ως προς τις τιμές της παραμέτρου μίξης μ, για τα δίκτυα LFR.



Σχήμα 7.2: Αποτελέσματα της τρίτης γεννήτριας τυχαίων περιπάτων (S_{any}), για διαφορετικές μετρικές ομοιότητας, ως προς τις τιμές της παραμέτρου μίξης μ, για τα δίκτυα LFR.



Σχήμα 7.3: Αξιολόγηση της απόδοσης των διαφορετικών διαδικασιών εξερεύνησης (και των συνδυασμών τους), όταν χρησιμοποιείται ο δείκτης Jaccard ως μετρική ομοιότητας, για τα τρία σύνολα δεδομένων του πραγματικού κόσμου, ως προς διαφορετικά ποσοστά δεδομένων εκπαίδευσης.

7.3.2 Σύνολα δεδομένων πραγματικού κόσμου

7.3.2.1 Σύγκριση των στρατηγικών ενσωμάτωσης

Το δεύτερο μέρος των πειραμάτων περιλαμβάνει τα σύνολα δεδομένων πραγματικού κόσμου. Το σχήμα 7.3 δείχνει τα αποτελέσματα των διαφορετικών διαδικασιών εξερεύνησης, καθώς και τους συνδυασμούς τους, όταν χρησιμοποιείται ο δείκτης Jaccard ως μετρική ομοιότητας, σε σύγκριση με τους δύο αλγόριθμους (DeepWalk και node2vec). Ο άξονας x αντιπροσωπεύει τα διαφορετικά ποσοστά δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση.

Οι προτεινόμενες στρατηγικές δίνουν διαφορετικά αποτελέσματα για κάθε σύνολο δεδομένων. Για τους λόγους που αναλύονται στην ενότητα 7.1 αναμένεται ότι, σε κάποιο βαθμό, οι επισημάνσεις του συνόλου δεδομένων BlogCatalog ευθυγραμμίζονται με τη δομή της κοινότητας του δικτύου. Για το λόγο αυτό αναμένεται συμπεριφορά παρόμοια με αυτή των τεχνητών συνόλων δεδομένων. Αυτό απεικονίζεται στο Σχήμα 7.3, όπου, για άλλη μια φορά, η στρατηγική S_{any}

έχει παρόμοιες επιδόσεις με τον αλγόριθμο DeepWalk. Αυτό είναι αναμενόμενο, καθώς η τρίτη γεννήτρια δεν λαμβάνει υπόψη τις συνδέσεις μεταξύ των κόμβων, οι οποίες είναι σημαντικές σε ένα δίκτυο με ισχυρές σχέσεις ομοφιλίας. Από την άλλη πλευρά, οι παραστάσεις που παράγονται από ένα συνδυασμό δύο ή περισσοτέρων διαφορετικών διαδικασιών εξερεύνησης έχουν απόδοση αντίστοιχη ή καλύτερη σε σχέση με το node2vec.

Στην περίπτωση του δικτύου Protein-Protein Interaction (PPI), οι ετικέτες δεν αντιπροσωπεύουν μόνο τις κοινότητες, αλλά επηρεάζονται επίσης από τα δομικά χαρακτηριστικά των πρωτεϊνών. Αυτό έχει σαν αποτέλεσμα να περιορίζεται η απόδοση όλων των μεθοδολογιών. Οι Grover et al. το τόνισαν αυτό επίσης στη δημοσίευσή τους, σημειώνοντας ότι τα αποτελέσματα της προσέγγισής τους (δηλαδή του node2vec) δεν διαφέρουν από αυτά του DeepWalk [38]. Αυτή η τάση επιβεβαιώνεται στην παρούσα μελέτη, όπου όλα τα σενάρια έχουν παρόμοια ή χειρότερα αποτελέσματα με το DeepWalk και το node2vec (σχήμα 7.3). Οι στρατηγικές με τις καλύτερες επιδόσεις φαίνεται να είναι αυτές που περιλαμβάνουν την τρίτη γεννήτρια τυχαίων περιπάτων (δηλαδή οι S_{any} και $U + S_{any}$). Καθώς το δίκτυο αυτό έχει ισοδυναμίες ρόλων μεταξύ των κόμβων, αναμένεται ότι οι στρατηγικές που βασίζονται στις ομοιότητες μεταξύ των κόμβων θα έχουν καλύτερες επιδόσεις.

Το τελικό πείραμα περιλαμβάνει το σύνολο δεδομένων Wikipedia, το οποίο, όπως αναφέρεται στην ενότητα 7.1, παρουσιάζει σε ένα βαθμό ισοδυναμίες δομής/ρόλου ανάμεσα στους κόμβους. Συνεπώς, αναμένεται μειωμένη απόδοση για αυτό το σύνολο δεδομένων, όπως φαίνεται και από το node2vec, το οποίο υπερβαίνει το DeepWalk με μικρό περιθώριο. Όσο για την προτεινόμενη μεθοδολογία, οι στρατηγικές που αποτελούνται από συνδυασμούς διαδικασιών εξερεύνησης επιτυγχάνουν και πάλι καλύτερα αποτελέσματα από τις ανεξάρτητες στρατηγικές (ειδικά η $U + S_{any}$).

Η στρατηγική με τις καλύτερες επιδόσεις συνολικά φαίνεται να είναι ο συνδυασμός της πρώτης και τρίτης γεννήτριας (δηλαδή η $U + S_{any}$). Αυτό επιβεβαιώνει την αρχική ιδέα, ότι οι ομοιότητες μεταξύ των κόμβων είναι εξίσου σημαντικές με τις συνδέσεις μεταξύ τους, ενώ η χρήση των πληροφοριών από τα δύο χαρακτηριστικά μπορεί να οδηγήσει σε καλύτερες αναπαραστάσεις.

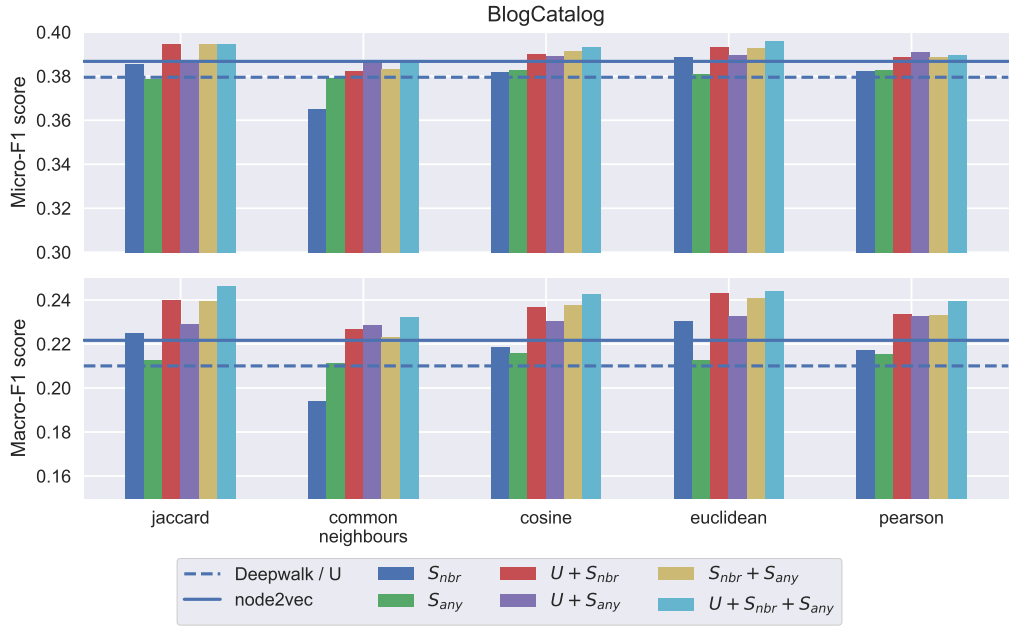
Ένα άλλο ενδιαφέρον συμπέρασμα είναι ότι οι πληροφορίες σχετικά με τις συνδέσεις και τις ομοιότητες πρέπει να έχουν ως αποτέλεσμα διαφορετικά χαρακτηριστικά (διαφορετικές στήλες στον τελικό πίνακα ενσωμάτωσης), ώστε ο ταξινομητής που χρησιμοποιεί αυτά τα χαρακτηριστικά να μπορεί να καθορίσει ποια από αυτά εφαρμόζονται καλύτερα στο εκάστοτε πρόβλημα.

Είναι επίσης σημαντικό να σημειωθεί ότι η στρατηγική που συνδυάζει όλες τις διαδικασίες εξερεύνησης ($U + S_{nbr} + S_{any}$) δεν δίνει απαραίτητα τα καλύτερα αποτελέσματα. Αυτό συμβαίνει κυρίως επειδή οι προκύπτουσες αναπαραστάσεις περιλαμβάνουν επαναλαμβανόμενη πληροφορία, η οποία οδηγεί σε συσχετισμένα

(correlated) χαρακτηριστικά.

Μια τελευταία παρατήρηση είναι ότι η προτεινόμενη μεθοδολογία δεν ξεπερνά πάντα τους αλγόριθμους με οποίους συγκρίνεται. Ωστόσο, είναι σημαντικό να δηλωθεί για μια ακόμη φορά ότι η προτεινόμενη προσέγγιση είναι μια πλήρως μη επιβλεπόμενη προσέγγιση και αποδίδει εξίσου καλά, ακόμη και όταν συγκρίνεται με ημι-επιβλεπόμενες προσεγγίσεις.

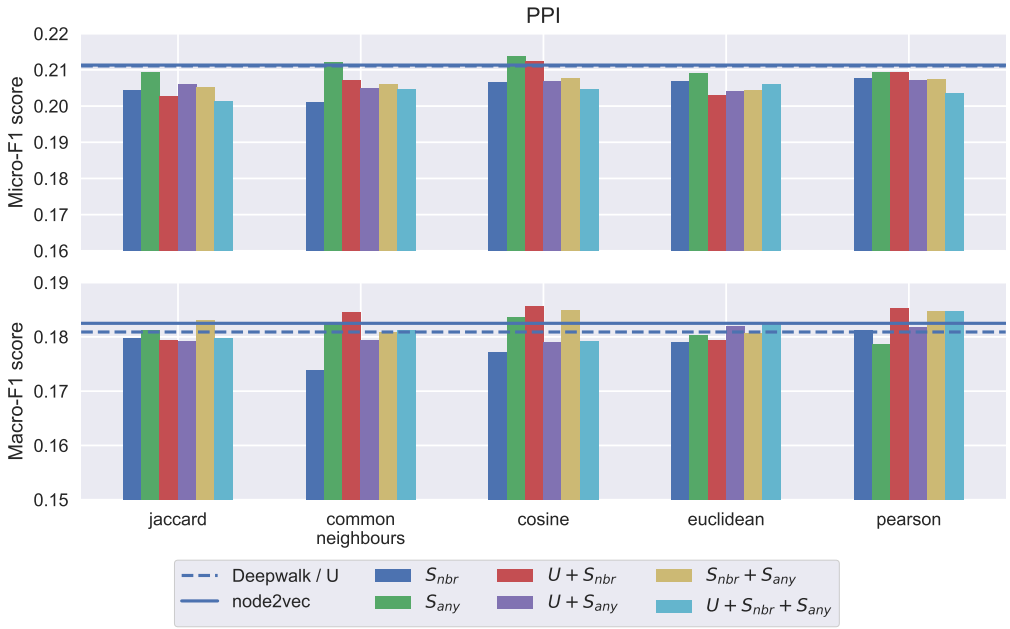
7.3.2.2 Σύγκριση των μετρικών ομοιότητας



Σχήμα 7.4: Η απόδοση των διαφορετικών στρατηγικών ενσωμάτωσης για το σύνολο δεδομένων BlogCatalog, όταν χρησιμοποιούνται διαφορετικές μετρικές ομοιότητας.

Μια κρίσιμη πτυχή της προτεινόμενης προσέγγισης είναι η επιλογή της μετρικής ομοιότητας που χρησιμοποιείται από την γεννήτρια τυχαίων περιπάτων. Για να το εξηγήσουμε καλύτερα, τα σχήματα 7.4, 7.5 και 7.6 απεικονίζουν την απόδοση για τα σύνολα δεδομένων BlogCatalog, Protein-Protein Interaction και Wikipedia, αντίστοιχα, όταν χρησιμοποιείται το 50% των κόμβων για εκπαίδευση. Χρησιμοποιείται η ίδια πειραματική διαδικασία, όπως και προηγουμένως. Τα DeepWalk και node2vec παρουσιάζονται ως οριζόντιες γραμμές κατά μήκος του σχήματος. Οι μπάρες αντιπροσωπεύουν τις διαφορετικές στρατηγικές ενσωμάτωσης, ενώ ο άξονας x αντιστοιχεί στις διαφορετικές μετρικές.

Το πιο σημαντικό συμπέρασμα είναι ότι η επιλογή της μετρικής ομοιότητας



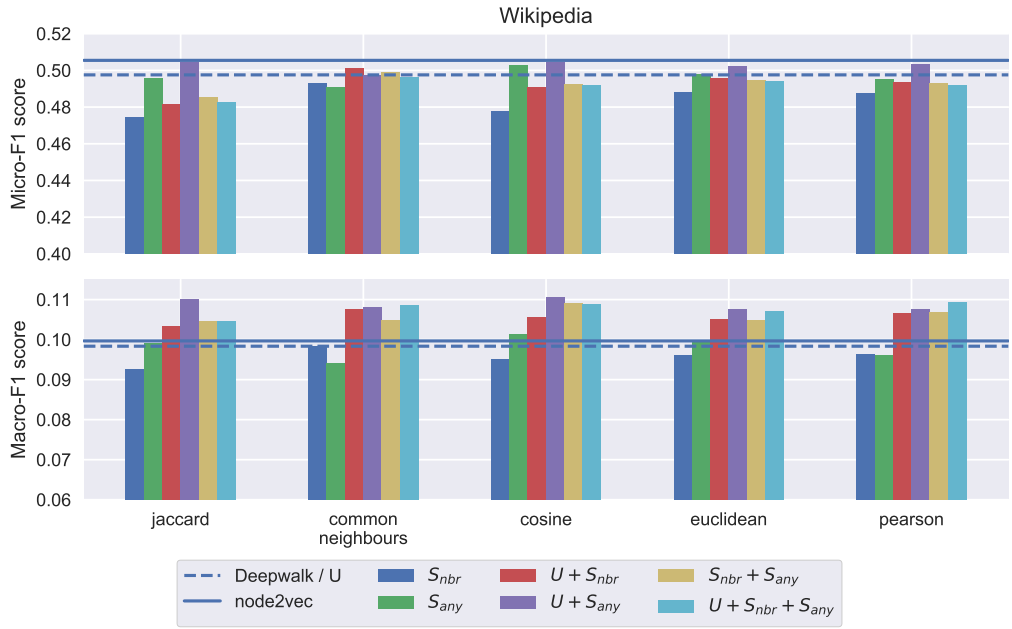
Σχήμα 7.5: Η απόδοση των διαφορετικών στρατηγικών ενσωμάτωσης για το σύνολο δεδομένων Protein-Protein Interaction (PPI), όταν χρησιμοποιούνται διαφορετικές μετρικές ομοιότητας.

δεν φαίνεται να επηρεάζει σημαντικά τα αποτελέσματα. Αυτό είναι μάλλον αναμενόμενο, καθώς πολλές από τις μετρικές είναι παρόμοιες μεταξύ τους. Το πιο αξιοσημείωτο παράδειγμα είναι ο δείκτης Jaccard, ο οποίος προκύπτει από τη μετρική των κοινών γειτόνων όταν συμπεριλαμβάνεται ένας συντελεστής κανονικοποίησης.

Η μόνη εξαίρεση φαίνεται να είναι η μετρική των κοινών γειτόνων, η οποία παρουσιάζει φτωχότερα αποτελέσματα για το σύνολο δεδομένων BlogCatalog, όπως φαίνεται στο σχήμα 7.4. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι αυτή η μετρική δεν έχει έναν συντελεστή κανονικοποίησης, επομένως μπορεί να ευνοεί κόμβους με περισσότερους γείτονες. Για τις υπόλοιπες μετρικές ομοιότητας, όλες οι συνδυαστικές στρατηγικές έχουν εξίσου καλά αποτελέσματα.

Στο σχήμα 7.5, εμφανίζονται τα αποτελέσματα για το σύνολο δεδομένων Protein-Protein Interaction. Η αρχική παρατήρηση ότι οι στρατηγικές S_{any} και $U + S_{any}$ έχουν τα καλύτερα αποτελέσματα συνεχίζει να ισχύει, ανεξάρτητα από την επιλογή της μετρικής ομοιότητας. Όπως εξηγήθηκε προηγουμένως (στην ενότητα 7.1), αυτό συμβαίνει λόγω των ισοδυναμιών ρόλων μεταξύ των κόμβων του συνόλου δεδομένων. Μια άλλη στρατηγική που επίσης δίνει καλά αποτελέσματα είναι η στρατηγική $U + S_{nbr}$.

Για το σύνολο δεδομένων Wikipedia (σχήμα 7.6), η στρατηγική $U + S_{any}$



Σχήμα 7.6: Η απόδοση των διαφορετικών στρατηγικών ενσωμάτωσης για το σύνολο δεδομένων Wikipedia, όταν χρησιμοποιούνται διαφορετικές μετρικές ομοιότητας.

εξακολουθεί να δίνει τα καλύτερα αποτελέσματα για όλες τις μετρικές ομοιότητας. Συνολικά, αποδεικνύεται και πάλι ότι το $U + S_{any}$ είναι η στρατηγική με τις καλύτερες επιδόσεις.

Τα αποτελέσματα που παρουσιάζονται στα σχήματα 7.4, 7.5 και 7.6 αποδεικνύουν ότι η επιλογή της μετρικής ομοιότητας δεν επηρεάζει σημαντικά την απόδοση, καθώς το σύστημα λειτουργεί εξίσου καλά ανεξάρτητα από την επιλεγμένη μετρική.

Κεφάλαιο 8

Ανίχνευση Λογοκλοπής Κειμένου

Ο συνεχώς αυξανόμενος όγκος πληροφοριών λόγω της χρήσης των υπολογιστών και του internet έχει αυξήσει την ανάγκη για αποτελεσματικές μεθόδους ανίχνευσης της λογοκλοπής. Η λογοκλοπή μπορεί να βρεθεί σε πολλά περιβάλλοντα με διαφορετική μορφή, για παράδειγμα στη λογοτεχνία, στις ακαδημαϊκές εργασίες και δημοσιεύσεις, ακόμα και στον κώδικα προγραμματισμού.

Η εγγενής ανίχνευση λογοκλοπής είναι το πεδίο έρευνας που ασχολείται με την εύρεση αντιγραμμένων αποσπασμάτων σε ένα έγγραφο, χωρίς την ύπαρξη εξωτερικών πηγών για σύγκριση με τα ύποπτα αποσπάσματα, αλλά εντοπίζοντας τις διαφοροποιήσεις στον τρόπο γραφής και τις ασυμφωνίες μέσα στο ίδιο το έγγραφο. Η βασική ιδέα είναι ότι μπορεί να δημιουργηθεί ένα προφίλ για τον τρόπο γραφής του συγγραφέα και να εντοπιστούν αποσπάσματα τα οποία διαφέρουν σημαντικά από αυτό.

8.1 Εισαγωγή

Η λογοκλοπή είναι η πράξη της αντιγραφής ή της μίμησης του έργου κάποιου άλλου και η παρουσίασή του ως πρωτότυπη, χωρίς όμως την κατάλληλη αναφορά ή παραπομπή. Η ανίχνευση λογοκλοπής σε έγγραφα κειμένου χωρίζεται σε δύο κύριες κατηγορίες, τις εξωγενείς και τις εγγενείς μεθόδους. Η διαφορά τους έγκειται στο αν απαιτείται, ή όχι, μια εξωτερική συλλογή εγγράφων για την ανίχνευση της λογοκλοπής.

Επομένως, οι εξωγενείς μέθοδοι συγκρίνουν μια συλλογή εγγράφων, η οποία αποτελεί πιθανή πηγή προέλευσης των αντιγραμμένων αποσπασμάτων, και ένα σύνολο ύποπτων εγγράφων, ενώ οι εγγενείς μέθοδοι προσδιορίζουν ποια από τα αποσπάσματα του εγγράφου υπό διερεύνηση είναι αντιγραμμένα, παρα-

τηρώντας τις διαφοροποιήσεις στον τρόπο γραφής μέσα στο ίδιο το κείμενο.

Η κεντρική ιδέα στην οποία βασίζεται η εγγενής ανίχνευση λογοκλοπής (Intrinsic Plagiarism Detection - IPD) είναι ότι κάθε συγγραφέας έχει το δικό του προσωπικό και μοναδικό στυλ γραφής, το οποίο μπορεί να ανιχνευθεί και να ποσοτικοποιηθεί χρησιμοποιώντας στυλιστικές ή/και σημασιολογικές τεχνικές.

Συνεπώς, αν αναλύσουμε ένα έγγραφο και αναζητήσουμε τα αποσπάσματα που δεν φαίνονται να ταιριάζουν με τον προσωπικό τρόπο γραφής του συγγραφέα, μπορούμε να ανιχνεύσουμε τη λογοκλοπή. Πρέπει βέβαια να ισχύει η εξής προϋπόθεση: το έγγραφο που εξετάζεται πρέπει να είναι γραμμένο κυρίως από τον ίδιο συγγραφέα, με μόνο μερικά κομμάτια κειμένου γραμμένα από άλλους συγγραφείς, έτσι ώστε να είναι εφικτό να εξαχθεί το στυλ γραφής του αρχικού συγγραφέα, ώστε να μπορεί στη συνέχεια να χρησιμεύσει ως σημείο αναφοράς.

Η εγγενής ανίχνευση λογοκλοπής είναι ένα πρόβλημα στενά συνδεδεμένο με άλλα ερευνητικά θέματα, όπως είναι η αναγνώριση συγγραφέα (author identification) και η επαλήθευση συγγραφέα (author verification), όπου, επίσης, η πιο σημαντική πηγή πληροφορίας είναι ο τρόπος γραφής του συγγραφέα.

Η αναγνώριση συγγραφέα [95], η οποία ονομάζεται επίσης και απόδοση συγγραφέα (author attribution), έχει ως στόχο να αποδώσει ένα έγγραφο ή ένα απόσπασμα σε έναν συγγραφέα με βάση το στυλ γραφής, μετρώντας τις στυλιστικές επιλογές μέσα στο κείμενο. Μπορεί να λάβει δύο μορφές: ο στόχος είναι είτε η ομαδοποίηση μιας συλλογής εγγράφων, έτσι ώστε κάθε ομάδα να αποτελείται από έγγραφα του ίδιου συγγραφέα, ή η ανίχνευση των συνόρων όπου αλλάζει ο συγγραφέας μέσα σε ένα έγγραφο το οποίο έχει προκύψει από συνεργασία. Η επαλήθευση συγγραφέα [53] απαντά στην ερώτηση αν ένα απόσπασμα γράφτηκε από έναν συγκεκριμένο συγγραφέα ή όχι, δεδομένης μιας συλλογής δειγμάτων κειμένου που είναι γνωστό ότι έχουν γραφτεί από τον συγκεκριμένο συγγραφέα.

Κάθε πρόβλημα αναγνώρισης συγγραφέα μπορεί να μετασχηματιστεί σε ένα σύνολο προβλημάτων επαλήθευσης συγγραφέα, με τον ίδιο τρόπο που κάθε πρόβλημα ταξινόμησης πολλαπλών κλάσεων (multi-class classification) μπορεί να μετατραπεί σε ένα σύνολο προβλημάτων ταξινόμησης μιας κλάσης (single-class classification) [96].

8.2 Θεωρητικό υπόβαθρο και σχετική έρευνα

8.2.1 Σχετική έρευνα

Μια ενδιαφέρουσα προσέγγιση του προβλήματος υλοποιήθηκε από τον Σταματάτο [94]. Το συγκεκριμένο σύστημα χρησιμοποιεί n-grams χαρακτήρων για

να αναγνωρίζει την αλλαγή του τρόπου γραφής, ενώ θεωρεί κάθε κείμενο ως σάκο από χαρακτήρες (bag of characters). Το προφίλ (Profile) του κειμένου είναι ένα διάνυσμα κανονικοποιημένων συχνοτήτων όλων των n-grams χαρακτήρων που εμφανίζονται στο κείμενο. Η διαφοροποίηση κάθε τμήματος κειμένου σε σχέση με ολόκληρο το έγγραφο υπολογίζεται χρησιμοποιώντας μια κανονικοποιημένη συνάρτηση απόστασης, η οποία επίσης προτείνεται από τον Σταματάτο [94]. Αυτή η συνάρτηση απόστασης ποσοτικοποιεί την ομοιότητα των προφίλ των n-grams όλων των τμημάτων σε σχέση με ολόκληρο το έγγραφο στο εύρος τιμών [0,1].

Ο Curran [23] παρουσιάζει μια προσέγγιση βασισμένη στα εξελικτικά νευρωνικά δίκτυα, με την ανάπτυξη ενός ταξινομητή εγγενούς ανίχνευσης λογοκλοπής, ο οποίος προσδιορίζει το στυλ γραφής του συγγραφέα ενός εγγράφου χάρη σε μια σειρά από στυλομετρικά χαρακτηριστικά.

Η μέθοδος ανίχνευσης λογοκλοπής των Oberreuter et al. [67, 68] κατασκευάζει μια σημασιολογική-λεξικολογική στυλιστική συνάρτηση, η οποία βασίζεται στη συχνότητα των όρων (term frequency). Η βασική ιδέα είναι η σύγκριση των συχνοτήτων των λέξεων μεταξύ κάθε τμήματος και ολόκληρου του εγγράφου. Για την ανίχνευση των ορίων ανάμεσα στα τμήματα κειμένου, το σύστημα χρησιμοποιεί ένα ad-hoc κατώφλι.

Οι Kestemont et al. [51] προτείνουν ένα σύστημα βασισμένο στα trigrams χαρακτήρων. Δημιούργησαν μια λίστα με τα trigrams χαρακτήρων που εμφανίζονται πιο συχνά στο σύνολο δεδομένων PAN 2009 και χρησιμοποίησαν αυτή τη λίστα για τη στυλιστική συνάρτηση τους, υπολογίζοντας τη συχνότητα εμφάνισης αυτών των trigrams σε κάθε τμήμα κειμένου. Για την ανίχνευση των ύποπτων αποσπασμάτων, σύγκριναν τα τμήματα εγγράφων μεταξύ τους, αντί να τα συγκρίνουν με ολόκληρο το έγγραφο.

Οι Ranatunga et al. [82] προτείνουν μια ενδιαφέρουσα προσέγγιση στην εγγενή ανίχνευση λογοκλοπής, δεδομένου ότι χρησιμοποιούν αυτο-οργανούμενους χάρτες (Self-Organizing Maps - SOMs) για να ομαδοποιήσουν τμήματα κειμένου από δύο διαφορετικούς συγγραφείς. Στον χάρτη που προκύπτει, οι ομάδες τμημάτων κειμένου που είναι τοπολογικά κοντά θεωρείται ότι ανήκουν στον ίδιο συγγραφέα. Επομένως ένα έγγραφο χωρίς λογοκλοπή πρέπει να αντιπροσωπεύεται εξ΄ ολοκλήρου από μία συμπαγή ομάδα.

Οι Tschuggnall et al. [103] προτείνουν τον αλγόριθμο *Plag-Inn*. Σύμφωνα με την συγκεκριμένη προσέγγιση, το κείμενο χωρίζεται σε προτάσεις και για κάθε πρόταση δημιουργείται ένα δέντρο γραμματικής (grammar tree). Στο δέντρο γραμματικής, κάθε λέξη αντικαθίσταται από την κλάση του μέρους του λόγου στο οποίο ανήκει. Στη συνέχεια υπολογίζεται η απόσταση μεταξύ των γραμματικών δέντρων και αποθηκεύεται σε έναν τριγωνικό πίνακα απόστασης. Θέτοντας ένα προκαθορισμένο όριο δ, κάθε πρόταση με μέση απόσταση μεγαλύτερη από δ θεωρείται ότι έχει προκύψει από λογοκλοπή. Αν και η ιδέα

πίσω από αυτόν τον αλγόριθμο είναι πολύ ενδιαφέρουσα, φαίνεται να είναι μη ρεαλιστική, αν λάβουμε υπ΄ όψιν τον υπολογιστικό χρόνο που απαιτείται για την εφαρμογή του σε ένα τυπικό έγγραφο με εκατοντάδες προτάσεις. Σε μια μεταγενέστερη εργασία [104], οι ίδιοι συγγραφείς προτείνουν μια προηγμένη έκδοση του αλγορίθμου, που ονομάζεται *PQ-PlagInn*.

Η μέθοδος των Hua et al. [48] χρησιμοποιεί τρεις διαφορετικές κατηγορίες χαρακτηριστικών, σε επίπεδο χαρακτήρων, λεξικού και μέρους του λόγου και χρησιμοποιεί ένα κατώφλι ως κριτήριο για την ανίχνευση της λογοκλοπής. Οι Alsallal et al. [3] παρουσιάζουν μια ολοκληρωμένη προσέγγιση που βασίζεται στη Λανθάνουσα Σημασιολογική Δεικτοδότηση (Latent Semantic Indexing - LSI) και τη στυλομετρία για την ανίχνευση λογοκλοπής. Το LSI χρησιμοποιείται για τον πίνακα των όρων των εγγράφων του συνόλου δεδομένων, ενώ η στυλομετρία χρησιμοποιείται για την προσέγγιση του τρόπου γραφής. Παρουσιάζουν αποτελέσματα στο σύνολο δεδομένων MED, μια συλλογή ιατρικών περιλήψεων.

Οι Bensalem et. al. [7] προτείνουν τις κλάσεις n-gram, οι οποίες υποδεικνύουν τη σχετική συχνότητα ενός n-gram μέσα στο έγγραφο, σε σχέση με το συχνότερο n-gram, και εφαρμόζουν μια μη επιβλεπόμενη μέθοδο ταξινόμησης που βασίζεται στον αλγόριθμο Naive Bayes.

Οι Kuta and Kitowski [54] προσπαθούν να βελτιώσουν την απόδοση της μεθόδου των προφίλ από n-grams χαρακτήρων που προτάθηκε από το Σταματάτο, ρυθμίζοντας τις παραμέτρους του και προτείνοντας τροποποιήσεις και πλούσια σύνολα χαρακτηριστικών. Ισχυρίζονται ότι επιτυγχάνουν καλύτερα αποτελέσματα από το Σταματάτο στο PAN 2009 και 2011, αλλά όπως αναφέρουν, δεν χωρίζουν τα δεδομένα σε σύνολα εκπαίδευσης (train set) και ελέγχου (test set).

Οι Kuznetsov et al. [55] προτείνουν μια μέθοδο η οποία χωρίζει κάθε κείμενο σε προτάσεις και κατασκευάζει βασικά στυλομετρικά χαρακτηριστικά για κάθε πρόταση (συχνότητες n-gram χαρακτήρων και λέξεων, αριθμός σημείων στίξης και αντωνυμιών) και έπειτα εκπαιδεύει έναν ταξινομητή (Gradient Boosting Regression Trees) χρησιμοποιώντας το σύνολο δεδομένων PAN 2011. Στη συνέχεια χρησιμοποιούν ένα κατώφλι, και όσες προτάσεις δίνουν τιμή υψηλότερη στην έξοδο του ταξινομητή, θεωρείται ότι έχουν προκύψει από λογοκλοπή.

8.2.2 Διαγωνισμοί PAN, σύνολα δεδομένων και συμμετέχοντα συστήματα.

Ο διαγωνισμός του PAN[1] είναι μια σειρά από επιστημονικά γεγονότα σχετικά με την εγκληματολογία ψηφιακού κειμένου. Μέχρι σήμερα, η εγγενής ανίχνευση λογοκλοπής έχει αποτελέσει αντικείμενο των διαγωνισμών PAN Webis τρεις φορές (τις χρονιές 2009, 2011 και 2016). Για κάθε διαγωνισμό δημιουργήθηκε ένα νέο σύνολο δεδομένων από τους διοργανωτές. Σε αυτή την ενότητα θα περιγράψουμε κάθε ένα από αυτά τα σύνολα δεδομένων, παρουσιάζοντας περιληπτικά τα πιο αποτελεσματικά συστήματα εγγενούς ανίχνευσης λογοκλοπής για κάθε σύνολο δεδομένων.

8.2.2.1 Διαγωνισμός PAN 2009.

Στο πλαίσιο του διαγωνισμού PAN 2009, δόθηκε στη διάθεση των ερευνητών ένα τεράστιο σώμα κειμένων το οποίο περιέχει «τεχνητή λογοκλοπή». Το σώμα κειμένων βασίζεται σε έγγραφα του Project Gutenberg [1, 77] και αποτελείται από 3092 έγγραφα για την εγγενή ανίχνευση λογοκλοπής. Τα μισά από αυτά περιέχουν περιπτώσεις τεχνητής λογοκλοπής, ενώ το ποσοστό των τμημάτων που έχουν προκύψει απο λογοκλοπή για ολόκληρο το σύνολο δεδομένων είναι περίπου 10%.

Το σύστημα που νίκησε σε αυτόν το διαγωνισμό υλοποιήθηκε από τον Σταματάτο [94]. Το σύστημα των Oberreuter et. al. [67], το οποίο ήταν το νικητήριο σύστημα του επόμενου διαγωνισμού εγγενούς ανίχνευσης λογοκλοπής, έδωσε καλύτερα αποτελέσματα σε αυτό το σύνολο δεδομένων. Επίσης, το μη συμμετέχον σύστημα των Bensalem et. al. [7], αξιολογήθηκε χρησιμοποιώντας αυτό το σύνολο δεδομένων και πέτυχε καλύτερα αποτελέσματα από το σύστημα του Σταματάτου.

8.2.2.2 Διαγωνισμός PAN 2011.

Το σώμα κειμένων για το συγκεκριμένο διαγωνισμό αποτελείται από 4753 έγγραφα. Περίπου το 6% των συνολικών προτάσεων στο σύνολο δεδομένων είναι προϊόν λογοκλοπής. Αυτό το σύνολο δεδομένων έχει μια αδυναμία ως προς το σχεδιασμό του: οι περιπτώσεις τεχνητής λογοκλοπής που εισάγονται στα ύποπτα έγγραφα επιλέγονται χωρίς να λαμβάνεται υπ΄ όψιν αν η θεματολογία των αρχικών εγγράφων συμπίπτει με τη θεματολογία του ύποπτου εγγράφου. Αυτό έχει ως αποτέλεσμα να έχουν εισαχθεί νέες λέξεις στα έγγραφα, οι οποίες δεν υπήρχαν εκ των προτέρων. Επομένως, το σημασιολογικό περιεχόμενο των

[1] http://pan.webis.de/

αντιγραμμένων αποσπασμάτων δεν έχει σχέση με αυτό του αρχικού κειμένου, το οποίο δεν είναι ρεαλιστικό για τις πραγματικές περιπτώσεις λογοκλοπής.

Το σύστημα των Oberreuter et. al. [67] είναι το σύστημα με τις καλύτερες συνολικά επιδόσεις στον διαγωνισμό PAN 2011. Παρά το γεγονός ότι το σύστημα αυτό επιτυγχάνει αξιοσημείωτα αποτελέσματα, η απόδοσή του θα πρέπει να αντιμετωπίζεται με σκεπτικισμό, καθώς θεωρείται ότι επωφελήθηκε από την αδυναμία του συνόλου δεδομένων, χρησιμοποιώντας χαρακτηριστικά βασισμένα στο λεξιλόγιο και ως εκ τούτου δεν θα ήταν σε θέση να γενικεύσει καλά σε ένα διαφορετικό, πιο ουδέτερο σύνολο δεδομένων [76, 4]. Ένα άλλο σύστημα που συμμετείχε στον διαγωνισμό PAN 2011 και κατέλαβε τη δεύτερη θέση ήταν το σύστημα των Kestemont et al [51].

Ορισμένα άλλα συστήματα εγγενούς ανίχνευσης λογοκλοπής, τα οποία δεν συμμετείχαν στους διαγωνισμούς, χρησιμοποίησαν αυτό το σύνολο δεδομένων για αξιολόγηση. Συγκεκριμένα, οι Tschuggnall et. al. [103] πειραματίστηκαν με 7 μόνο έγγραφα του συνόλου δεδομένων του PAN 2011, γεγονός που δεν αποτελεί αξιόπιστη αξιολόγηση. Οι ίδιοι συγγραφείς [104] χρησιμοποίησαν μια μεταγενέστερη έκδοση του αλγορίθμου τους σε ένα μικρό υποσύνολο του σώματος κειμένων του PAN 2011 (50 έγγραφα). Τέλος, το σύστημα των Bensalem et. al. [7] εφαρμόζεται και σε αυτό το σύνολο δεδομένων.

8.2.2.3 Διαγωνισμός PAN 2016.

Στα πλαίσια του διαγωνισμού PAN 2016 η εγγενής ανίχνευση λογοκλοπής συμπεριλήφθηκε ως υπο-πρόβλημα της Διαφοροποίησης Συγγραφέα (Author Diarization) [86]. Ο στόχος της διαφοροποίησης συγγραφέα είναι ο προσδιορισμός διαφορετικών συγγραφέων μέσα σε ένα ενιαίο έγγραφο, ομαδοποιώντας τμήματα του κειμένου που ανήκουν στον ίδιο συγγραφέα.

Ο αριθμός διαφορετικών συγγραφέων, και κατ΄ επέκταση ο αριθμός των ομάδων, μπορεί να είναι, ή όχι, γνωστός εκ των προτέρων. Στην ειδική περίπτωση, όπου μπορούμε να υποθέσουμε ότι το μεγαλύτερο κομμάτι του κειμένου είναι γραμμένο από έναν συγγραφέα και μόνο μερικά τμήματα γράφονται από άλλους συγγραφείς, η διαφοροποίηση συγγραφέα συμπίπτει με την εγγενή ανίχνευση λογοκλοπής.

Το σύνολο δεδομένων του PAN 2016 για την εγγενή ανίχνευση λογοκλοπής είναι μικρό, καθώς αποτελείται από μόλις 71 έγγραφα. Επιπλέον, το μέσο μήκος των εγγράφων είναι πολύ μικρότερο σε σχέση με τα προηγούμενα έτη. Όλα τα έγγραφα περιέχουν περιπτώσεις τεχνητής λογοκλοπής, ενώ το ποσοστό της λογοκλοπής είναι περίπου 13%. Υπάρχουν μόνο δύο συμμετέχοντες, και τα αποτελέσματά τους είναι μάλλον φτωχά σε σύγκριση με τα αποτελέσματα των προηγούμενων ετών. Η πρώτη θέση του διαγωνισμού PAN 2016 ανήκει στους Kuznetsov et al.[55]. Τα υποδεέστερα αποτελέσματα μπορεί να αποδο-

θούν στο μικρό μέγεθος των εγγράφων στο συγκεκριμένο σύνολο δεδομένων, που δεν επαρκεί για την εξαγωγή πληροφορίας σε σχέση με τον τρόπο γραφής του συγγραφέα.

Στον παρακάτω πίνακα, μπορούμε να δούμε τα στατιστικά στοιχεία για τα τρία σύνολα δεδομένων. Είναι προφανές ότι σε όλα τα σύνολα δεδομένων υπάρχει ανισορροπία στον αριθμό των λογοκλεμμένων και μη λογοκλεμμένων προτάσεων. Τα μη ισορροπημένα σύνολα δεδομένων είναι ένα σημαντικό ζήτημα κατά την εκπαίδευση μοντέλων μηχανικής μάθησης[101], καθώς μπορούν να επηρεάσουν την απόδοση του συστήματος.

	PAN'09	*PAN'11*	*PAN'16*
Αριθμός εγγράφων	3092	4753	71
Μέσο μέγεθος εγγράφου (λέξεις)	43,057	33,155	1,678
Μέσο μέγεθος εγγράφου (προτάσεις)	1,962	1,396	73
Έγγραφα που περιέχουν λογοκλοπή (%)	50	49.8	100
Λογοκλεμμένες λέξεις (%)	9.52	5.9	12.62
Λογοκλεμμένες προτάσεις (%)	9.37	5.53	12.96

8.2.3 Καθιερωμένη μεθοδολογία

Γενικά, τα συστήματα εγγενούς ανίχνευσης λογοκλοπής βασίζονται σε μία τυπική μεθοδολογία τριών βημάτων: 1) κατάτμηση του κειμένου, 2) ανάλυση του στυλ και 3) αναγνώριση των έκτοπων σημείων. Ωστόσο, εκτός από αυτά τα βήματα, μπορεί να συμπεριληφθούν επιπλέον βήματα προεπεξεργασίας και μετεπεξεργασίας.

8.2.3.1 Βήμα 1 - Κατάτμηση του κειμένου.

Όπως εξηγήθηκε προηγουμένως, κάθε συγγραφέας έχει το δικό του στυλ γραφής. Σε ένα λογοκλεμμένο έγγραφο, είναι αναμενόμενο το στυλ γραφής να μεταβάλλεται από το ένα τμήμα του κειμένου στο άλλο. Επομένως, το πρώτο βήμα ενός συστήματος εγγενούς ανίχνευσης λογοκλοπής είναι να χωρίσει το έγγραφο σε τμήματα, τα οποία στη συνέχεια συγκρίνονται είτε με ολόκληρο το έγγραφο είτε με τα γειτονικά τμήματα. Αυτά τα τμήματα μπορεί να είναι φυσικά τμήματα του εγγράφου, όπως προτάσεις, παράγραφοι ή κεφάλαια [28, 98], ή μπορεί να είναι μπλοκ κειμένου καθορισμένου μήκους (σε χαρακτήρες) [94, 67].

Η πιο συνηθισμένη μέθοδος κατάτμησης χρησιμοποιεί ένα *μετακινούμενο παράθυρο*, το οποίο κινείται κατά έναν αριθμό χαρακτήρων ή προτάσεων σε κάθε βήμα επεξεργασίας. Τα περιεχόμενα του παραθύρου για κάθε βήμα δίνονται, στη συνέχεια, ως είσοδος στη συνάρτηση ανάλυσης στυλ. Οι δύο παράμετροι

αυτής της διαδικασίας είναι το *μήκος παραθύρου* και το *βήμα παραθύρου*. Η επιλογή του μήκους του παραθύρου είναι ιδιαίτερα σημαντική. Εάν το μήκος του παραθύρου είναι πολύ μικρό, ενδέχεται να μην καταφέρνει να καταγράφει τις στυλιστικές ιδιότητες του κειμένου, ενώ ένα μεγαλύτερο παράθυρο μπορεί να κατατάξει σε λάθος κατηγορία τμήματα μικρότερου μεγέθους.

8.2.3.2 Βήμα 2 - Ανάλυση του στυλ.

Στο δεύτερο βήμα ενός συστήματος εγγενούς ανίχνευσης λογοκλοπής, τα τμήματα κειμένου δίνονται ως είσοδος στη συνάρτηση ανάλυσης στυλ, η οποία εξάγει το στυλιστικό αποτύπωμα για κάθε ένα από αυτά (δηλαδή τις τιμές για ένα σύνολο χαρακτηριστικών που σχετίζονται με το στυλ γραφής). Η συνάρτηση ανάλυσης στυλ είναι κατασκευασμένη με βάση μια σειρά από στυλομετρικά ή/και σημασιολογικά χαρακτηριστικά. Η στυλομετρία (stylometry) χρησιμοποιεί στατιστικές μεθόδους για την ανάλυση του λογοτεχνικού στυλ [47], ενώ τα σημασιολογικά χαρακτηριστικά εξάγουν πληροφορίες για τον πλούτο του λεξιλογίου και το σημασιολογικό πλαίσιο του αποσπάσματος [20]. Οι συγγραφείς του [98] παρέχουν μια συλλογή από τα πιο γνωστά στυλομετρικά χαρακτηριστικά, τα οποία μπορεί να είναι λεξικολογικά, συντακτικά ή δομικά.

Για την εξαγωγή της στυλιστικής υπογραφής του αρχικού συγγραφέα, υπάρχουν δύο επιλογές: είτε εφαρμόζεται η συνάρτηση ανάλυσης στυλ σε ολόκληρο το έγγραφο, είτε παράγεται από το συνδυασμό της πληροφορίας από τα επιμέρους τμήματα. Η έξοδος της συνάρτησης ανάλυσης στυλ είναι ένα διάνυσμα που περιέχει τα προαναφερθέντα χαρακτηριστικά, τα οποία παίρνουν πραγματικές τιμές. Το διάνυσμα αυτό αξιολογείται αργότερα για να προσδιοριστεί αν το συγκεκριμένο τμήμα εγγράφου περιέχει ή όχι αντιγραμμένο κείμενο.

8.2.3.3 Βήμα 3 - Αναγνώριση των έκτοπων σημείων.

Ο στόχος του τελικού βήματος της εγγενούς ανίχνευσης λογοκλοπής είναι να ανιχνεύσει τα έκτοπα σημεία, κατηγοριοποιώντας τα αποσπάσματα του κειμένου σε λογοκλεμμένα και μη.

Υπάρχουν δύο διαφορετικές επιλογές για το πως θα γίνει η σύγκριση στο συγκεκριμένο βήμα. Είτε συγκρίνεται το στυλιστικό αποτύπωμα κάθε τμήματος του εγγράφου με την εκτιμώμενη υπογραφή του αρχικού συγγραφέα , είτε συγκρίνονται τα στυλιστικά αποτυπώματα των τμημάτων εγγράφων μεταξύ τους. Η υπογραφή του συγγραφέα εξάγεται εφαρμόζοντας τη συνάρτηση ανάλυσης στυλ σε ολόκληρο το έγγραφο ή στο σώμα κειμένου που αποδίδεται στο δημιουργό.

Για τη σύγκριση επιλέγεται και εφαρμόζεται μια κατάλληλη συνάρτηση απόστασης *(π.χ. συντελεστής συνημιτόνου, απόσταση Manhattan)* [4] στα διανύ-

σματα χαρακτηριστικών που προέκυψαν από το προηγούμενο βήμα. Το αποτέλεσμα αυτής της διαδικασίας είναι μια μοναδική τιμή, η οποία αντιστοιχεί στη στυλιστική απόσταση μεταξύ των δύο τμημάτων κειμένου.

Στη συνέχεια, για να επισημανθούν τα όρια ανάμεσα στα τμήματα, ορίζεται ένα κατώφλι για την τιμή της απόστασης και τα τμήματα κειμένου που εμφανίζουν τιμές πάνω από αυτό το κατώφλι θεωρούνται έκτοπα σημεία. Η τιμή του κατώτατου ορίου μπορεί να επιλεγεί είτε με ad hoc τρόπο, όπως στα περισσότερα υπάρχοντα συστήματα, από πρότερη γνώση και παρατήρηση των δεδομένων εκπαίδευσης [94, 104], ή με τη χρήση μεθόδων μηχανικής μάθησης στα δεδομένα εκπαίδευσης [28, 23, 48, 7].

8.3 Περιγραφή του συστήματος

Σε αυτή την ενότητα, παρουσιάζουμε μια προσέγγιση εγγενούς ανίχνευσης λογοκλοπής, η οποία βασίζεται σε μεθόδους μηχανικής μάθησης. Θα περιγράψουμε το σύστημα ανίχνευσης λογοκλοπής και θα αναλύσουμε τις επιλογές μοντελοποίησης που έγιναν. Το σύστημα αποτελείται από ένα βήμα προεπεξεργασίας, από τα τρία βασικά τμήματα που αναλύθηκαν στο προηγούμενο κεφάλαιο (τμηματοποίηση κειμένου, ανάλυση στυλ - εξαγωγή χαρακτηριστικών, ανίχνευση έκτοπων σημείων) και ένα βήμα μετεπεξεργασίας των αποτελεσμάτων.

Η προτεινόμενη προσέγγιση αποτελεί επέκταση συστήματος το οποίο υλοποιήθηκε στα πλαίσια εκπόνησης διπλωματικής εργασίας του εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης, το οποίο περιγράφεται στα [36, 75].

8.3.1 Προεπεξεργασία

Η προεπεξεργασία αποτελείται από τα εξής βήματα:

- Μετατροπή κεφαλαίων γραμμάτων σε πεζά
- Ανίχνευση προτάσεων
- Ανίχνευση λεκτικών μονάδων
- Αφαίρεση αλφαριθμητικών και ειδικών χαρακτήρων
- Επισήμανση των λέξεων ως μερών του λόγου
- Αναγνώριση θεμάτων των λέξεων

Πίνακας 8.1: Τιμές παραμέτρων για τη μέθοδο μετακινούμενου παραθύρου ανάλογα με το μέγεθος κειμένου

	Μέγεθος κειμένου (s σε kB)	**Μέγεθος παραθύρου (σε προτάσεις)**	**Μέγεθος βήματος (σε προτάσεις)**
Μικρό	$s < 30$	12	4
Μεσαίο	$30 \leq s < 300$	15	5
Μεγάλο	$300 \leq s$	30	10

8.3.2 Κατάτμηση κειμένου

Για την κατάτμηση του κειμένου εφαρμόζουμε τη μέθοδο του *μετακινούμενου παραθύρου* με δυο διαφορετικούς τρόπους:

- Στην πρώτη περίπτωση χρησιμοποιούμε σταθερές τιμές για τις δύο παραμέτρους, οι οποίες είναι 15 προτάσεις για το *μέγεθος παραθύρου* και 5 προτάσεις για την *τιμή βήματος*.

- Στη δεύτερη περίπτωση, πειραματιζόμαστε με διαφορετικές τιμές παραμέτρων, οι οποίες αλλάζουν ανάλογα με το μέγεθος του κειμένου υπό εξέταση. Συγκεκριμένα χωρίζουμε τα κείμενα σε μικρού, μεσαίου και μεγάλου μεγέθους, και ανάλογα χρησιμοποιούμε τις τιμές των παραμέτρων που δίνονται στον πίνακα 8.1.

Η ιδέα πίσω από τις μεταβαλλόμενες τιμές παραμέτρων είναι ότι ένα μικρό έγγραφο αναμένεται να έχει μικρότερα τμήματα που έχουν προκύψει από λογοκλοπή, σε σχέση με ένα πολύ μεγάλο κείμενο, και το αντίστροφο.

Όπως εξηγήθηκε και νωρίτερα, η επιλογή του μεγέθους του παραθύρου είναι κρίσιμη και επηρεάζει άμεσα την απόδοση των συστημάτων ανίχνευσης λογοκλοπής. Η επιλογή μικρού μεγέθους παραθύρου σημαίνει καλύτερη προσαρμογή στα λογοκλεμμένα τμήματα του εγγράφου, όμως μπορεί να μην δίνει αρκετή πληροφορία στη συνάρτηση στυλιστικής ανάλυσης για την εξαγωγή αξιόπιστων στατιστικών στοιχείων. Από την άλλη, ένα μεγαλύτερο παράθυρο μπορεί να κατατάξει σε λάθος κατηγορία τα τμήματα του κειμένου με μέγεθος μικρότερο του παραθύρου, εφόσον στο εσωτερικό του παραθύρου θα συνυπάρχει μη λογοκλεμμένο και λογοκλεμμένο κείμενο.

Επομένως, ένα μετακινούμενο παράθυρο που μεταβάλλει τις παραμέτρους του ώστε να προσαρμόζεται ακριβώς στα αποσπάσματα που έχουν προκύψει από λογοκλοπή, είναι πιο πιθανό να επιστρέφει τμήματα κειμένου που περιέχουν υψηλό ποσοστό λογοκλεμμένου κειμένου.

8.3.3 Στυλιστική ανάλυση - Εξαγωγή χαρακτηριστικών

Για τη στυλιστική ανάλυση εξάγουμε 11 χαρακτηριστικά, κάποια στυλιστικά και κάποια σημασιολογικά. Στην λίστα που ακολουθεί, τα χαρακτηριστικά 1-4 χρησιμοποιούνται από τα περισσότερα υπάρχοντα συστήματα [98, 20], τα 5-7 είναι νέα στυλιστικά χαρακτηριστικά που βασίζονται στην ιδέα της συμπίεσης και τα 8-12 είναι σημασιολογικά χαρακτηριστικά που βασίζονται στην έννοια της κλάσης συχνότητας λέξης (word frequency class):

1. μέσο μήκος πρότασης (average sentence length)
2. μέσο πλήθος συλλαβών ανά λέξη (average syllable count per token)
3. έλεγχος αναγνωσιμότητας του Flesch (Flesch reading-ease test) [27]
4. συχνότητα της λέξης "of"
5. ποσοστό συμπίεσης ρημάτων (verbs' compression rate)
6. ποσοστό συμπίεσης επιρρημάτων (adverbs' compression rate)
7. ποσοστό συμπίεσης επιθέτων (adjectives' compression rate)
8. μέσος όρος των θετικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου - χωρίου (mean value of positive subtraction of word frequency class between document - segment - wfc_1)
9. τυπική απόκλιση των θετικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου-χωρίου (standard deviation of positive subtraction of word frequency class between document - segment - wfc_2)
10. ποσοστό λέξεων που εμφανίζουν θετική διαφορά κλάσης συχνότητας μεταξύ κειμένου-χωρίου πάνω από το μέσο όρο (percentage of the words having word frequency class subtraction value between document - segment greater than the mean value - wfc_3)
11. μέσος όρος των αρνητικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου- χωρίου για τις συχνά εμφανιζόμενες[2], σε όλο το κείμενο, λέξεις (mean value of negative subtraction of word frequency class between document - segment, only for the frequent words - wfc_4)

[2]μετά από πειράματα, υποδηλώνουμε μια λέξη ως *συχνή* αν η τιμή συχνότητας λέξης κλάσης είναι μικρότερη από 1,8, σε σχέση με την εξίσωση 8.2

8.3.3.1 Γνωστά στυλιστικά χαρακτηριστικά

Τα πρώτα τέσσερα από τα χαρακτηριστικά που εφαρμόζουμε είναι γνωστά στον τομέα της υπολογιστικής γλωσσολογίας. Έχοντας χωρίσει το ύποπτο έγγραφο σε προτάσεις, μπορούμε να μετρήσουμε το *μέσο μήκος πρότασης*, σε χαρακτήρες, τόσο για κάθε κομμάτι του εγγράφου, όσο και για ολόκληρο το έγγραφο. Με τον εντοπισμό των λέξεων του ύποπτου εγγράφου μπορούμε να μετρήσουμε το *μέσο όρο συλλαβών ανά λέξη* που εμφανίζεται στο τμήμα ενός εγγράφου. Ο έλεγχος αναγνωσιμότητας του Flesch έχει σχεδιαστεί για να υποδεικνύει πόσο δύσκολο είναι να κατανοηθεί ένα απόσπασμα στα αγγλικά και δίνεται από την ακόλουθη εξίσωση:

$$\mathrm{F} = 206.835 - 1.015\left(\frac{\text{λέξεις}}{\text{προτάσεις}}\right) - 84.6\left(\frac{\text{συλλαβές}}{\text{λέξεις}}\right) \tag{8.1}$$

8.3.3.2 Νέα χαρακτηριστικά βασισμένα στη συμπίεση

Τα νέα στυλιστικά χαρακτηριστικά *(ποσοστό συμπίεσης ρημάτων, επιρρημάτων, επιθέτων)* βασίζονται στην ιδέα των Leanne και Matwin [92], οι οποίοι πρότειναν στυλιστικές μετρικές σχετιζόμενες με την πολυπλοκότητα Kolmogorov (Kolmogorov Complexity). Η βασική ιδέα, είναι πως κάθε τμήμα ενός κειμένου παρουσιάζει μια κατανομή για κάθε κλάση λέξεων. Με τη χρήση ενός απλού αλγόριθμου που προσπαθεί να εκφράσει την πιθανή κατανομή μέσω του ποσοστού συμπίεσης, για κάθε μια από τις τρεις κλάσεις λέξεων με τις οποίες έχουμε πειραματιστεί, κάνουμε την εξαγωγή μιας δυαδικής ακολουθίας, όπου οι λέξεις που ανήκουν στην συγκεκριμένη κλάση λέξεων αντιπροσωπεύονται με 1, ενώ εκείνες που δεν ανήκουν αντιπροσωπεύονται με 0. Στη συνέχεια, η δυαδική ακολουθία συμπιέζεται χρησιμοποιώντας τον αλγόριθμο *κωδικοποίησης μήκους διαδρομής* (run length encoding). Τέλος, το ποσοστό συμπίεσης δίνεται από την εξίσωση 8.2.

$$\text{Ποσοστό συμπίεσης} = \frac{\text{Μήκος κωδικοποίησης Run-length}}{\text{Μήκος δυαδικής ακολουθίας}} \tag{8.2}$$

8.3.3.3 Σημασιολογικά χαρακτηριστικά βασισμένα στην κλάση συχνότητας λέξης

Ονομάζουμε τα χαρακτηριστικά $(8-11)$ σημασιολογικά, αφού θεωρούμε ότι τα χαρακτηριστικά που σχετίζονται με τον πλούτο του λεξιλογίου μπορούν να θεωρηθούν ως τέτοια. Αυτά τα χαρακτηριστικά είναι βασισμένα στο θέμα (stem) της λέξης, το οποίο υποδεικνύει τη σημασία των λέξεων. Επομένως, τα πιο συχνά χρησιμοποιούμενα θέματα λέξεων σε ένα έγγραφο υποδεικνύουν

το γενικό πλαίσιο στο οποίο αναφέρεται το έγγραφο. Από εδώ και πέρα, όταν αναφερόμαστε σε *λέξεις* σε σχέση με την κλάση συχνότητας λέξης εννοούμε *θέματα λέξεων* (stems).

Τα σημασιολογικά χαρακτηριστικά (8−11) βασίζονται στην έννοια της *κλάσης συχνότητας λέξης*, όπως περιγράφεται από τους Meyer zu Eissen και Stein [28].

Αν C είναι ένα σύνολο δεδομένων και $f(w)$ είναι η συχνότητα της λέξης w στο C, τότε η κλάση συχνότητας λέξης $c(w)$ της λέξης w υπολογίζεται από την εξίσωση 8.3, όπου w^* είναι η πιο συχνή λέξη στο C.

$$c(w) = log_2 \left(\frac{f(w^*)}{f(w)} \right) \tag{8.3}$$

Πιο αναλυτικά, η πιο συνηθισμένη λέξη παίρνει μηδενική τιμή κλάσης συχνότητας λέξης, ενώ όσο πιο σπάνια είναι μια λέξη, τόσο υψηλότερη θα είναι η τιμή της κλάσης συχνότητας λέξης. Αξίζει να σημειωθεί ότι η κλάση συχνότητας λέξης εξαρτάται από το κομμάτι του εγγράφου στο οποίο εφαρμόζεται, αφού σχετίζεται με τη συνηθέστερη λέξη.

Κατά την στυλιστική ανάλυση, υπολογίσαμε την κλάση συχνότητας λέξης των λέξεων σε ολόκληρο το έγγραφο αλλά και για κάθε κομμάτι του εγγράφου ξεχωριστά.

Τα χαρακτηριστικά αυτά έχουν ως τελικό στόχο τη σύγκριση των τιμών κλάσης συχνότητας λέξης των λέξεων που εμφανίζονται σε κάθε τμήμα του εγγράφου με τις αντίστοιχες τιμές τους σε ολόκληρο το έγγραφο. Η ιδέα είναι ότι συχνές λέξεις σε ένα απόσπασμα κειμένου, που όμως είναι σπάνιες σε ολόκληρο το έγγραφο, αποτελούν ένδειξη πιθανής λογοκλοπής σε αυτό το απόσπασμα. Αυτό αντικατοπτρίζεται από μια υψηλή τιμή στη διαφορά μεταξύ της κλάσης συχνότητας λέξης του περάσματος και του εγγράφου. Συγκεκριμένα, η θετική τιμή της διαφοράς υποδεικνύει ότι οι πολύ συχνές λέξεις (θέματα λέξεων) στο απόσπασμα παρουσιάζονται σπάνια στο έγγραφο ως σύνολο. Το αντίθετο ισχύει για την αρνητική τιμή της διαφοράς.

8.3.3.4 Κανονικοποίηση χαρακτηριστικών

Το τελευταίο βήμα της εξαγωγής χαρακτηριστικών για κάθε απόσπασμα είναι η σύγκριση με ολόκληρο το έγγραφο και, τέλος, η κανονικοποίηση των τιμών. Τα σημασιολογικά χαρακτηριστικά σχετίζονται με όλο το έγγραφο εξ΄ ορισμού. Για τα στυλιστικά χαρακτηριστικά $(1-7)$ αφαιρούμε τις τιμές κάθε αποσπάσματος από την αντίστοιχη τιμή όλου του εγγράφου. Τέλος, κανονικοποιούμε τις τιμές των χαρακτηριστικών εφαρμόζοντας την εξίσωση 8.4.

$$\text{normalise}(x) = \frac{x}{1 + |x|} \tag{8.4}$$

Όπως αναφέρθηκε προηγουμένως, σε κάθε τμήμα εγγράφου αντιστοιχεί ένα διάνυσμα χαρακτηριστικών. Κάθε στοιχείο του διανύσματος περιέχει την απόσταση του αποσπάσματος από το έγγραφο για το συγκεκριμένο χαρακτηριστικό.

Συγκεκριμένα, έστω f ένα χαρακτηριστικό, f_s η τιμή του για ένα τμήμα και f_d η τιμή του για ολόκληρο το έγγραφο. Η τιμή που προκύπτει για το συγκεκριμένο τμήμα, πριν από την κανονικοποίηση, είναι ίση με $(f_d - f_s)$, και οι τιμές αυτές βρίσκονται σε μεγάλο και ακανόνιστο εύρος. Το σημαντικό είναι ότι αυτή η τιμή της απόστασης αναμένεται να είναι κοντά στο 0 για τα μη λογοκλεμμένα τμήματα, ενώ παίρνει υψηλότερη τιμή στην περίπτωση των λογοκλεμμένων τμημάτων.

Οι ταξινομητές μας χρειάζονται ένα κανονικό φάσμα τιμών για την εκπαίδευση τους. Συγκεκριμένα, χρειαζόμαστε τις τιμές εισόδου του τμήματος μηχανικής μάθησης να βρίσκονται στην περιοχή [0,1]. Επιπλέον, μας ενδιαφέρει πολύ να διατηρήσουμε τιμές που δεν αντιστοιχούν σε λογοκλοπή κοντά στο 0, και αυτές που αντιστοιχούν σε λογοκλοπή κοντά στο 1, αντίστοιχα.

Έτσι, χρειαζόμαστε μια συνάρτηση κανονικοποίησης με τις ακόλουθες ιδιότητες:

1. Να επιστρέφει τιμές στο διάστημα [0,1].

2. Να μην μεταβάλλει σημαντικά τις τιμές που βρίσκονται κοντά στο 0.

3. Να επιστρέφει τιμές κοντά στο 1 για τις υψηλότερες τιμές.

Κάθε συνάρτηση που έχει αυτές τις ιδιότητες θεωρείται κατάλληλη για το σύστημά μας. Κατά τη διάρκεια της έρευνάς μας πειραματιστήκαμε και με άλλες παρόμοιες συναρτήσεις. Καταλήξαμε στην εξίσωση 8.4 λόγω της απλότητας και της αποτελεσματικότητάς της.

8.3.4 Αναγνώριση των έκτοπων σημείων

Η υλοποίηση του τμήματος της αναγνώρισης των έκτοπων σημείων βασίζεται σε μεθόδους μηχανικής μάθησης. Σε αυτή την ενότητα αναφερόμαστε σε αυτές τις μεθόδους, περιγράφουμε τον τρόπο εκπαίδευσής τους και εξηγούμε τον τρόπο με τον οποίο αντιμετωπίζουμε το πρόβλημα μη ισορροπημένων δεδομένων το οποίο εμφανίζεται στο σύνολο δεδομένων μας.

8.3.4.1 Ταξινόμηση με τη χρήση μηχανικής μάθησης

Πειραματιζόμαστε με τρεις ταξινομητές: τα Δένδρα Αποφάσεων (Decision Trees), τις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) και τα Τυχαία Δάση (Random Forests).

Για τη δημιουργία του συνόλου εκπαίδευσης αναθέτουμε επισημάνσεις στα διανύσματα χαρακτηριστικών με τον εξής τρόπο: αν τουλάχιστον το 50% ενός αποσπάσματος έχει προκύψει από λογοκλοπή, τότε όλο το απόσπασμα θεωρείται ότι έχει προκύψει από λογοκλοπή. Τα διανύσματα χαρακτηριστικών σε συνδυασμό με τις αντίστοιχες επισημάνσεις δίνονται ως είσοδος σε κάθε ταξινομητή κατά την εκπαίδευση.

Στη συνέχεια, ο ταξινομητής παίρνει ως είσοδο ένα διαφορετικό σύνολο διανυσμάτων χαρακτηριστικών, για τα οποία δεν υπάρχουν επισημάνσεις, και τα οποία πρέπει να κατατάξει σε λογοκλεμμένα ή μη λογοκλεμμένα. Παρ΄ ότι το τμήμα μηχανικής μάθησης του συστήματος εκπαιδεύεται και προβλέπει σε επίπεδο τμήματος κειμένου, τα τελικά αποτελέσματα του συστήματος είναι σε επίπεδο πρότασης, λόγω ενός βήματος μετα-επεξεργασίας, το οποίο περιγράφεται αργότερα με λεπτομέρεια.

8.3.4.2 Εξισορρόπηση του συνόλου δεδομένων

Από τη φύση της, η εγγενής ανίχνευση λογοκλοπής είναι μια διαδικασία που διαχειρίζεται μη ισορροπημένα δεδομένα, αφού τα τμήματα κειμένου που έχουν προκύψει από λογοκλοπή τείνουν να είναι λιγότερα σε αριθμό, με μεγάλη διαφορά, από τα πρωτότυπα τμήματα. Αυτό το γεγονός επηρεάζει την απόδοση του συστήματος, επομένως είναι κρίσιμο να αντιμετωπιστεί. Για το λόγο αυτό εφαρμόζουμε τεχνικές εξισορρόπησης (balancing techniques) στα δεδομένα μας.

Για την εξισορρόπηση των δεδομένων εκπαίδευσης χρησιμοποιούμε και συγκρίνουμε τα αποτελέσματα ενός αριθμού διαφορετικών μεθόδων. Αρχικά, η εξισορρόπηση των δεδομένων γίνεται με τη χρήση του αλγόριθμου SMOTE-borderline [12]. Ο αλγόριθμος SMOTE-borderline, ένας αλγόριθμος υπερδειγματοληψίας (oversampling), δημιουργεί συνθετικά παραδείγματα της κλάσης μειονότητας χρησιμοποιώντας τις τιμές των ήδη υπαρχόντων παραδειγμάτων της κλάσης μειονότητας και των γειτονικών τους, υπό την προϋπόθεση ότι και αυτά ανήκουν στην κλάση μειονότητας.

Η δεύτερη προσέγγισή μας είναι να πειραματιστούμε με μια μέθοδο που συνδυάζει τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας (undersampling), η οποία ονομάζεται SMOTE + ENN [5]. Η μέθοδος SMOTE + ENN είναι μια μέθοδος εξισορρόπησης που προκύπτει από την εφαρμογή του αλγορίθμου SMOTE-borderline που ακολουθείται από τον αλγόριθμο Επεξεργασμένων

Πλησιέστερων Γειτόνων (Edited Nearest Neighbors - ENN) [112].

Η τεχνική ENN είναι μια τεχνική υποδειγματοληψίας, η οποία αφαιρεί κάθε παράδειγμα του οποίου η κλάση διαφέρει από την κλάση της πλειοψηφίας των κοντινότερών του γειτόνων. Με άλλα λόγια, «επεξεργάζεται» το σύνολο δεδομένων, αφαιρώντας δείγματα που δεν συμφωνούν «αρκετά» με τη γειτονιά τους, επομένως επηρεάζει τα δείγματα κοντά στα όρια των κλάσεων και, έτσι, μπορεί να εξαλείψει την αλληλεπικάλυψη ανάμεσα στις κλάσεις.

Το κίνητρο πίσω από αυτή την προσέγγιση είναι ότι ο SMOTE μπορεί να παράγει θορυβώδη δείγματα στα όρια των κλάσεων, παρεμβάλλοντας νέα σημεία μεταξύ δειγμάτων που αποκλίνουν απο την υπόλοιπη κλάση. Το ζήτημα αυτό μπορεί να λυθεί με τον καθαρισμό του χώρου που λαμβάνεται μετά από την υπερδειγματοληψία, χρησιμοποιώντας τον ENN.

Τέλος, πειραματιζόμαστε με το πρώτο βήμα της μεθόδου SMOTE + ENN, εφαρμόζοντας διαφορετικά ποσοστά κατά την υπερδειγματοληψία. Αντί της δειγματοληψίας με στόχο την πλήρη ισορροπία ανάμεσα στις δύο κλάσεις, αυξάνουμε το μέγεθος της τάξης μειονότητας σε ένα ποσοστό του μεγέθους της τάξης πλειοψηφίας. Η σκέψη πίσω από αυτή την επιλογή είναι ότι, στο σύνολο δεδομένων μας, ο λόγος των λογοκλεμμένων αποσπασμάτων, σε σχέση με τα μη λογοκλεμμένα, είναι μικρός, και σε κάποιες περιπτώσεις μπορεί να φτάσει το 1:20. Αυτό σημαίνει ότι η χρήση μιας μεθόδου υπερδειγματοληψίας με στόχο την εξίσωση των μεγεθών των δύο κλάσεων πρακτικά διπλασιάζει το μέγεθος του συνόλου δεδομένων μας, πράγμα που αυξάνει την πιθανότητα υπερπροσαρμογής (overfitting) για την κλάση μειονότητας, ενώ επιβραδύνει τη διαδικασία εκπαίδευσης.

8.3.4.3 Μετεπεξεργασία

Το τελευταίο τμήμα του συστήματος έχει ως στόχο την τελική επεξεργασία των προβλέψεων για κάθε τμήμα του εγγράφου, όπως αυτές έχουν δοθεί από το προηγούμενο βήμα της μηχανικής μάθησης. Επεξεργαζόμαστε αυτά τα αποτελέσματα για δύο λόγους. Όπως περιγράψαμε προηγουμένως, η εφαρμογή της μεθόδου του μετακινούμενου παραθύρου με τιμή βήματος k και μέγεθος παραθύρου $3k$, οδηγεί σε επικαλυπτόμενα παράθυρα, αφού κάθε ομάδα k προτάσεων περιλαμβάνεται σε 3 διαδοχικά παράθυρα. Συνεπώς, έχουμε 3 διαφορετικές προβλέψεις για κάθε τέτοια ομάδα προτάσεων (αν δεν λάβουμε υπ΄ όψιν την πρώτη και την τελευταία ομάδα προτάσεων). Εφαρμόζοντας τη διαδικασία μετεπεξεργασίας που περιγράφεται παρακάτω, εξάγουμε την τελική πρόβλεψη για κάθε πρόταση, αφού επισημαίνουμε τελικά κάθε ομάδα k προτάσεων, επομένως και κάθε πρόταση, μόνο μία φορά.

Ο δεύτερος λόγος που κάνουμε αυτή την μετεπεξεργασία είναι ώστε τα αποτελέσματά μας να είναι συγκρίσιμα με τα συστήματα που συμμετέχουν στους

διαγωνισμούς του PAN. Οι συμμετέχοντες οφείλουν να παρέχουν τα αποτελέσματα τους σε επίπεδο χαρακτήρα, πράγμα που όμως είναι ενάντια στη διαίσθηση και την πραγματικότητα. Μια τέτοια πρακτική θεωρεί περιπτώσεις λογοκλοπής που ξεκινούν ή τελειώνουν στη μέση μιας φράσης ή μιας λέξης, το οποίο είναι μη ρεαλιστικό. Έτσι, αποφασίσαμε να αξιολογήσουμε το σύστημά μας με βάση τα στατιστικά στοιχεία σε επίπεδο προτάσεων, τα οποία θεωρούμε εύλογα ως προς τις πραγματικές περιπτώσεις λογοκλοπής και τα οποία προσεγγίζουν περισσότερο τους συμμετέχοντες του PAN, σε σχέση με τα στατιστικά στοιχεία σε επίπεδο τμημάτων κειμένου.

Δεδομένων των προβλέψεων για κάθε τμήμα του εγγράφου, όπως αυτές έχουν προκύψει από τους ταξινομητές, η διαδικασία μετεπεξεργασίας έχει ως εξής. Για κάθε ομάδα k προτάσεων, υπολογίζουμε τη μέση τιμή mv των επισημάνσεων που αντιστοιχούν στα τρία επικαλυπτόμενα παράθυρα που περιέχουν αυτή την ομάδα. Εάν $mv \geq \frac{2}{3}$, τότε η ομάδα προτάσεων επισημαίνεται ως λογοκλεμμένη. Διαφορετικά, χαρακτηρίζεται ως μη λογοκλεμμένη. Είναι σχεδιαστική επιλογή μας να θεωρήσουμε μια ομάδα φράσεων ως λογοκλεμμένες αν τουλάχιστον 2 από τα 3 τμήματα που τις περιέχουν προβλέπονται ως λογοκλεμμένα. Κατά συνέπεια, κάθε ομάδα k προτάσεων θα λάβει την ίδια επισήμανση πρόβλεψης. Ωστόσο, κάθε πρόταση έχει τη δική της επισήμανση για να μπορούμε να εξάγουμε τα στατιστικά στοιχεία σε επίπεδο πρότασης, όπως αναφέρθηκε παραπάνω.

Κεφάλαιο 9

Ανίχνευση Λογοκλοπής Κειμένου - Πειραματικό μέρος

9.1 Αξιολόγηση

Σε αυτό το κεφάλαιο παρουσιάζεται μια σειρά πειραμάτων με στόχο την αξιολόγηση του προτεινόμενου συστήματος. Πειραματιζόμαστε με όλες τις παραμέτρους του συστήματος που συζητήθηκαν μέχρι στιγμής. Πιο συγκεκριμένα, για το βήμα κατάτμησης κειμένου πειραματιζόμαστε τόσο με σταθερό όσο και με μεταβαλλόμενο μήκος παραθύρου. Για το βήμα μηχανικής μάθησης πειραματιζόμαστε με διαφορετικούς συνδυασμούς μεθόδων εξισορρόπησης δεδομένων και ταξινομητών.

Όπως εξηγήθηκε στο προηγούμενο κεφάλαιο, διαχωρίζουμε τα χαρακτηριστικά μας σε δύο υποκατηγορίες: στα στυλομετρικά χαρακτηριστικά *(μέσος μήκος πρότασης, μέσο πλήθος συλλαβών ανά λέξη, έλεγχος αναγνωσιμότητας του Flesch, συχνότητα της λέξης "of", ποσοστό συμπίεσης ρημάτων, επιρρημάτων και επιθέτων)* και στα σημασιολογικά χαρακτηριστικά *(τα τέσσερα χαρακτηριστικά με βάση την κλάση συχνότητας λέξης)*.

Προκειμένου να εκτιμηθεί η σημασία των διαφορετικών τύπων χαρακτηριστικών, δοκιμάζουμε διαφορετικά σύνολα χαρακτηριστικών στα σύνολα δεδομένων του PAN. Επομένως, παρέχουμε αποτελέσματα για τα ακόλουθα τρία σύνολα χαρακτηριστικών: *i*) για όλα τα χαρακτηριστικά, *ii*) μόνο για τα στυλομετρικά χαρακτηριστικά, *iii*) μόνο για τα σημασιολογικά χαρακτηριστικά.

Επιπλέον, παρέχουμε μια λίστα με το F-score το οποίο προκύπτει όταν δίνουμε σαν είσοδο κάθε στυλιστικό χαρακτηριστικό μόνο του, προκειμένου να αξιολογήσουμε τα νέα χαρακτηριστικά που προτείνουμε.

Τέλος, συγκρίνουμε τα καλύτερα αποτελέσματα του συστήματός μας για τους τρεις ταξινομητές με τα καλύτερα συστήματα των διαγωνισμών PAN.

Σε όλες τις περιπτώσεις εφαρμόζουμε τη μέθοδο k-fold cross validation με τιμές παραμέτρων που δίνονται στους πίνακες 9.1, 9.2.

Η εγγενής ανίχνευση λογοκλοπής είναι ένα πρόβλημα ταξινόμησης μιας κλάσης, όπου τα ψευδώς αρνητικά παίζουν σημαντικό ρόλο. Για το λόγο αυτό, αξιολογούμε το σύστημά μας χρησιμοποιώντας ως μετρικές την ακρίβεια, την ανάκληση και το μέτρο F1 (precision, recall, F1-score).

Πίνακας 9.1: 4-fold cross validation για το σύνολο δεδομένων του PAN 2009.

	Έγγραφα που περιλαμβάνονται	
	Σύνολο Εκπαίδευσης	**Σύνολο Δοκιμής**
split 1	$\{773, 3092\}$	$\{1, 772\}$
split 2	$\{1, 772\} \cup \{1547, 3092\}$	$\{773, 1546\}$
split 3	$\{1, 1546\} \cup \{2321, 3092\}$	$\{1547, 2320\}$
split 4	$\{1, 2320\}$	$\{2321, 3092\}$

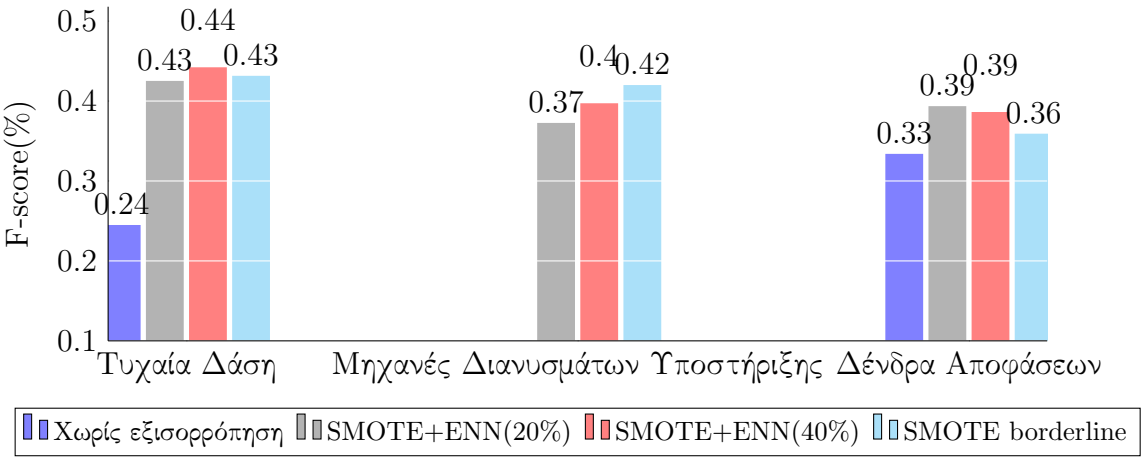
Πίνακας 9.2: 5-fold cross validation για το σύνολο δεδομένων του PAN 2011.

	Έγγραφα που περιλαμβάνονται	
	Σύνολο Εκπαίδευσης	**Σύνολο Δοκιμής**
split 1	$\{1001, 4753\}$	$\{1, 1000\}$
split 2	$\{1, 1000\} \cup \{2001, 4753\}$	$\{1001, 2000\}$
split 3	$\{1, 2000\} \cup \{3001, 4753\}$	$\{2001, 3000\}$
split 4	$\{1, 3000\} \cup \{4001, 4753\}$	$\{3001, 4000\}$
split 4	$\{1, 4000\}$	$\{4001, 4753\}$

9.1.1 Ταξινομητές και μέθοδοι εξισορρόπησης

Αρχικά συγκρίνουμε τη συμπεριφορά του συστήματός μας για διαφορετικούς συνδυασμούς ταξινομητών και μεθόδων εξισορρόπησης δεδομένων. Οι τρεις διαφορετικές παραλλαγές, σχετικά με τις μεθόδους εξισορρόπησης, που εφαρμόστηκαν στα δεδομένα εκπαίδευσης είναι οι SMOTE, SMOTE + ENN και SMOTE + ENN με διαφορετικά ποσοστά υπερδειγματοληψίας. Στην τρίτη προσέγγιση, πειραματιστήκαμε με διαφορετικά ποσοστά υπερδειγματοληψίας, αυξάνοντας το μέγεθος της κλάσης μειονότητας ώστε να γίνει ίσο με ένα ποσοστό του μεγέθους της τάξης πλειονότητας (20%, 40%, 60% και 80%) και παρουσιάζουμε τις καλύτερες επιδόσεις. Για αυτό το πείραμα θεωρούμε σταθερό μήκος παραθύρου 15 προτάσεων για το τμήμα κατάτμησης κειμένου.

Οι εικόνες 9.1 και 9.2 απεικονίζουν τα αποτελέσματα που αντιστοιχούν στα σύνολα δεδομένων PAN 2009 και PAN 2011, αντίστοιχα. Παρατηρούμε ότι οι



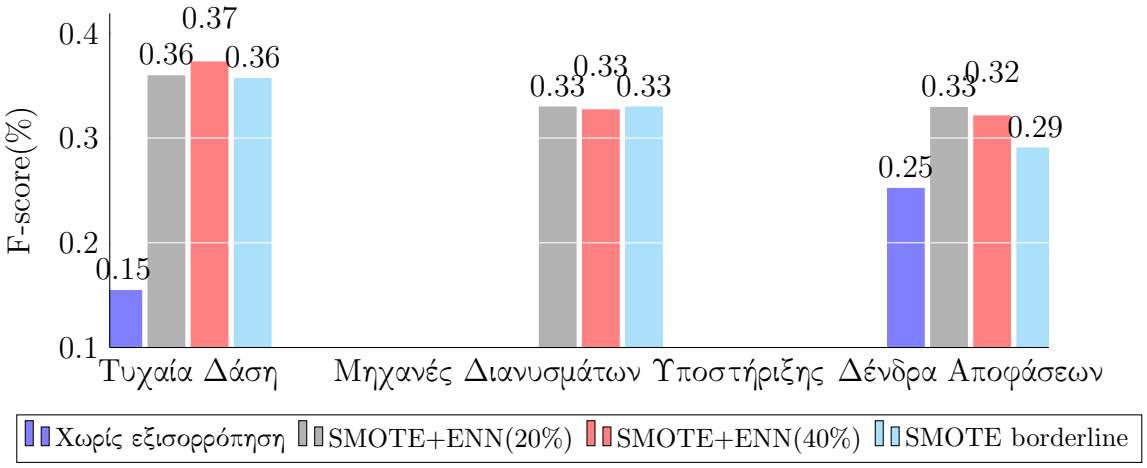
Σχήμα 9.1: Αποτελέσματα για το σύνολο δεδομένων του PAN 2009

μηχανές διανυσμάτων υποστήριξης δεν ανταποκρίνονται καλά στα μη ισορροπημένα δεδομένα εκπαίδευσης. Ο ταξινομητής δεν είναι ικανός να μάθει από τα δείγματα της κλάσης μειονότητας. Επιπλέον, τα τυχαία δάση και τα δέντρα αποφάσεων δίνουν πολύ χειρότερα αποτελέσματα όταν τα δεδομένα εκπαίδευσης δεν είναι ισορροπημένα. Όσο για τις μεθόδους εξισορρόπησης, όλες συμβάλλουν στην καλύτερη απόδοση σε οποιονδήποτε ταξινομητή. Ωστόσο, φαίνεται ότι δεν υπάρχει προφανής νικητής μεταξύ αυτών, καθώς κάθε ταξινομητής συμπεριφέρεται καλύτερα για διαφορετική μέθοδο εξισορρόπησης. Τα τυχαία δάση είναι πιο αποτελεσματικά όταν εφαρμόζεται η συνδυασμένη μέθοδος εξισορρόπησης, ενώ οι μηχανές διανυσμάτων υποστήριξης συνδυάζονται καλύτερα με την τεχνική SMOTE borderline, ενώ το SMOTE + ENN είναι επίσης προτιμότερο για τα δέντρα αποφάσεων αλλά με μικρότερη αναλογία υπερδειγματοληψίας. Συνολικά, ο συνδυασμός τυχαίο δάσος με SMOTE + ENN επιτυγχάνει τα καλύτερα αποτελέσματα.

9.1.2 Κατάτμηση κειμένου

Σε ότι παρουσιάσαμε μέχρι στιγμής έχουμε χρησιμοποιήσει σταθερό μήκος παραθύρου. Ο πίνακας 9.3 περιέχει τα αποτελέσματα της μεθόδου μετακινούμενου παραθύρου με μεταβαλλόμενο μήκος παραθύρου, το οποίο αλλάζει ανάλογα με το μέγεθος του εγγράφου, όπως συζητήθηκε στην ενότητα 8.3.2. Παρουσιάζουμε τα αποτελέσματα για τους καλύτερους συνδυασμούς μεθόδων εξισορρόπησης και ταξινομητών.

Τα αποτελέσματα δεν είναι ακριβώς αυτά που περιμέναμε. Αν και η μέθοδος συναγωνίζεται τη μέθοδο με σταθερό μήκος παραθύρου, δεν δίνει καλύτερα αποτελέσματα. Για όλους τους ταξινομητές έχουμε ελαφρώς χειρότερη απόδοση



Σχήμα 9.2: Αποτελέσματα για το σύνολο δεδομένων του PAN 2011

Πίνακας 9.3: Αποτελέσματα της μεθόδου μετακινούμενου παραθύρου με μεταβαλλόμενο μήκος παραθύρου

Σύνολο δεδομένων	**Ταξινομητής**	**Precision**	**Recall**	**F-score**
PAN 2009	RF	0.355	0.544	0.430
	SVM	0.340	0.526	0.413
	DT	0.345	0.450	0.390
PAN 2011	RF	0.338	0.375	0.356
	SVM	0.266	0.431	0.329
	DT	0.289	0.336	0.311
PAN 2016	SVM	0.126	0.352	0.186

(περίπου 1%).

Ωστόσο, η ιδέα είναι ότι ένα μεγαλύτερο έγγραφο είναι πιο πιθανό να περιέχει μεγαλύτερα λογοκλεμμένα αποσπάσματα, σε σχέση με ένα μικρότερο. Επομένως, μια μέθοδος κατάτμησης, που προσπαθεί να προσαρμόσει το μέγεθος των τμημάτων των εγγράφων ανάλογα με το συνολικό μέγεθός τους, αναμένεται να ταιριάζει καλύτερα στα λογοκλεμμένα τμήματα και έτσι να διευκολύνει τα επόμενα τμήματα του συστήματος να τα αναγνωρίσουν. Στην ιδανική περίπτωση, ο ρόλος μιας μεθόδου κατάτμησης είναι να επιστρέφει τμήματα που περιέχουν είτε καθαρά λογοκλεμμένο, είτε καθαρά μη λογοκλεμμένο κείμενο. Η μέθοδός μας δίνει ικανοποιητικά αποτελέσματα αλλά δεν ανταποκρίνεται στις προσδοκίες μας. Πιστεύουμε ότι οι τρεις διαφορετικοί συνδυασμοί των τιμών των παραμέτρων πιθανώς δεν ήταν βέλτιστες για τα σύνολα δεδομένων μας.

Όσον αφορά το σύνολο δεδομένων PAN 2016, το σύστημά μας παράγει μάλλον ανεπαρκή αποτελέσματα, τα οποία όμως είναι πολύ κοντά σε εκείνα του

καλυτέρου συστήματος από αυτά που συμμετείχαν στο διαγωνισμό [55] (σχεδόν 20% F-score). Όπως αναφέρθηκε ήδη, αυτό το σύνολο δεδομένων αποτελείται από ένα μικρό αριθμό σύντομων εγγράφων. Στην εγγενή ανίχνευση λογοκλοπής χρειαζόμαστε το ύποπτο έγγραφο να περιέχει μεγάλο αριθμό λέξεων, ώστε να παρέχει αρκετή πληροφορία για να μπορεί το σύστημα να εξάγει το πρωτότυπο στυλ γραφής. Επιπλέον, τα λογοκλεμμένα τμήματα των εγγράφων του PAN 2016 είναι τόσο σύντομα ώστε τα δέντρα αποφάσεων και τα τυχαία δάση δεν επιστρέφουν καθόλου τις περιπτώσεις λογοκλοπής (γι΄ αυτό τα αποτελέσματά τους δεν συμπεριλαμβάνονται στον πίνακα 9.3). Επιπλέον, στη βιβλιογραφία, μόνο δύο ομάδες συγγραφέων [55, 93], παρέχουν αποτελέσματα για την εγγενή ανίχνευση λογοκλοπής για το σύνολο δεδομένων του PAN 2016. Επομένως, πιστεύουμε ότι το σύνολο δεδομένων PAN 2016 δεν είναι κατάλληλο για το πρόβλημα και γι΄ αυτό το λόγο στα συγκριτικά πειραματικά αποτελέσματά μας που δίνονται στην ενότητα 9.1.5 συγκρίνουμε την προσέγγισή μας με άλλες μεθόδους μόνο στα σύνολα δεδομένων PAN 2009 και 2011.

9.1.3 Σταθερότητα του συστήματος ανίχνευσης

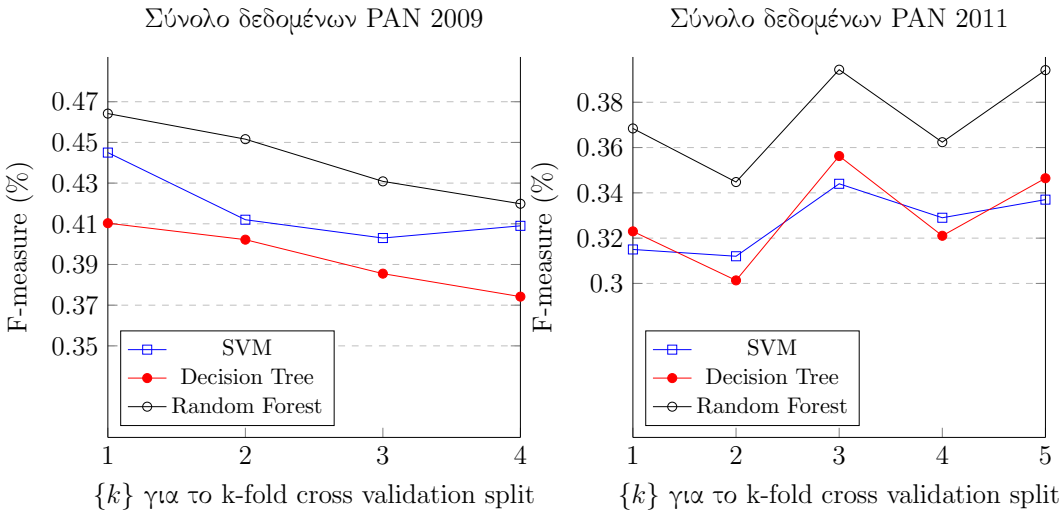
Σε αυτή την ενότητα ελέγχουμε τη σταθερότητα και την αξιοπιστία του συστήματός μας εξετάζοντας τα αποτελέσματα για κάθε split της μέθοδο k-fold cross validation. Προφανώς, ένα σταθερό σύστημα είναι πιο πιθανό να δώσει προβλέψεις με την αναμενόμενη ακρίβεια, ακόμη και όταν εφαρμόζεται σε ένα άγνωστο σύνολο δεδομένων, καθώς λειτουργεί με τον ίδιο τρόπο ακόμα και σε δύσκολες περιπτώσεις και ακραίες τιμές.

Όπως φαίνεται στο σχήμα 9.3 για τα δύο σύνολα δεδομένων, το σύστημά μας είναι σταθερό. Σε καμία περίπτωση δεν έχουμε απόκλιση μεγαλύτερη του 5% μεταξύ των αποτελεσμάτων των διαφορετικών splits. Επιπλέον, όταν εφαρμόζουμε εξισορρόπηση του συνόλου δεδομένων εκπαίδευσης το σύστημα τείνει να είναι πιο σταθερό.

9.1.4 Κατάταξη των χαρακτηριστικών

Στον πίνακα 9.4, παρουσιάζουμε τα αποτελέσματα για κάθε ένα από τα χαρακτηριστικά, για το σύνολο των εγγράφων του PAN 2009. Για αυτό το πείραμα χρησιμοποιήσαμε μηχανές διανυσμάτων υποστήριξης σε συνδυασμό με την τεχνική εξισορρόπησης δεδομένων SMOTE borderline.

Είναι προφανές ότι τα νέα στυλομετρικά χαρακτηριστικά *(συμπίεση ρημάτων, επιρρημάτων και επιθέτων)* είναι ιδιαίτερα ανταγωνιστικά σε σχέση με τα άλλα γνωστά και ευρέως χρησιμοποιούμενα χαρακτηριστικά. Μόνο το χαρακτηριστικό που αντιστοιχεί στο πλήθος συλλαβών ανά λέξη επιτυγχάνει υψηλότερη βαθμολογία, ενώ ο έλεγχος αναγνωσιμότητας του Flesch και το μέσο μήκος



Σχήμα 9.3: Αποτελέσματα για κάθε split των δύο συνόλων δεδομένων, για όλα τα χαρακτηριστικά

πρότασης αποδεικνύονται πολύ υποδεέστερα. Επιπλέον, τα σημασιολογικά χαρακτηριστικά που βασίζονται στην κλάση συχνότητας λέξης δίνουν εξαιρετικά ελπιδοφόρα αποτελέσματα.

Ωστόσο, η δύναμη του συστήματός μας έγκειται στην ποικιλία των εφαρμοσμένων χαρακτηριστικών και όχι σε ένα μοναδικό ισχυρό χαρακτηριστικό. Πολλά φαινομενικά ανεξάρτητα στυλιστικά χαρακτηριστικά, σε συνδυασμό με ένα κατάλληλο σύστημα μηχανικής μάθησης που είναι σε θέση να ανακαλύψει τις υποκείμενες σχέσεις μεταξύ τους, είναι πιο πιθανό να αναγνωρίσουν τις στυλιστικές διαφοροποιήσεις απ΄ ότι μερικά ισχυρά χαρακτηριστικά μόνα τους.

9.1.5 Σύγκριση με άλλα συστήματα

Οι εικόνες 9.4 και 9.5 παρουσιάζουν τα αποτελέσματα των καλύτερων συνδυασμών παραμέτρων για το σύστημά μας για τους τρεις ταξινομητές μαζί με τα νικητήρια συστήματα των διαγωνισμών PAN 2009, 2011, καθώς και τα συστήματα των Bensalem et al. και των Kuta και Kitowski. Τα τέσσερα συστήματα αξιολογήθηκαν και στα δύο σύνολα δεδομένων (2009 και 2011). Παρ΄ όλα αυτά, το σύστημα των Kuta και Kitowski δεν χωρίζει το σώμα κειμένων σε σύνολα εκπαίδευσης και ελέγχου. Επίσης, σημειώνεται ότι στην εικόνα 9.4, το σύστημα των Kuta και Kitowski έχει ελαφρώς υψηλότερη βαθμολογία σε σχέση με των Bensalem et al. (0.3341 και 0.33, αντίστοιχα).

Όπως αναφέρθηκε ήδη, τα σύνολα δεδομένων δεν είναι εντελώς ρεαλιστικά, αφού οι περιπτώσεις τεχνητής λογοκλοπής δημιουργούνται με την εισαγωγή

Πίνακας 9.4: Το F-score για κάθε χαρακτηριστικό στο σύνολο δεδομένων PAN 2009.

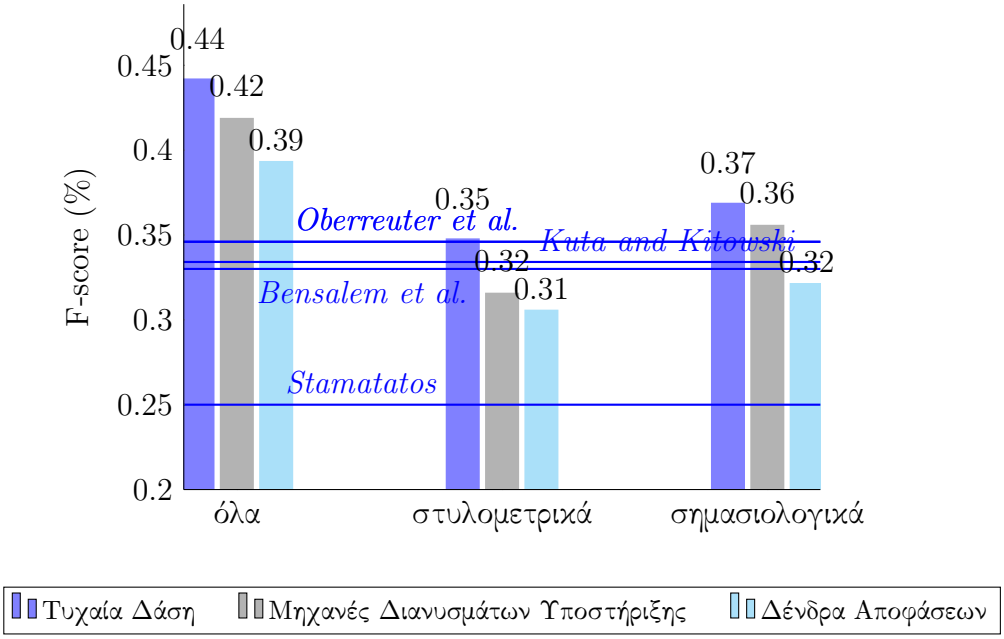
	Χαρακτηριστικό	**F-score**
Στυλομετρικά χαρακτηριστικά	μέσο πλήθος συλλαβών ανά λέξη	0.280
	ποσοστό συμπίεσης ρημάτων	0.276
	ποσοστό συμπίεσης επιθέτων	0.239
	ποσοστό συμπίεσης επιρρημάτων	0.233
	συχνότητα της λέξης "of"	0.226
	έλεγχος αναγνωσιμότητας του Flesch	0.199
	μέσο μήκος πρότασης	0.198
Σημασιολογικά χαρακτηριστικά	wfc_4	0.359
	wfc_3	0.225
	wfc_1	0.210
	wfc_2	0.184

αποσπασμάτων διαφορετικής θεματολογίας στα αρχικά κείμενα. Ως εκ τούτου, θεωρούμε σκόπιμο να ελέγξουμε χωριστά την απόδοση του συστήματος όταν παίρνει σαν είσοδο μόνο τα σχετικά αμερόληπτα στυλομετρικά χαρακτηριστικά, καθώς και όταν παίρνει σαν είσοδο μόνο τα σημασιολογικά. Αυτό συμβαίνει λόγω του ελαττωματικού σχεδιασμού των συνόλων δεδομένων, τα οποία κατασκευάστηκαν με εισαγωγή αποσπασμάτων άσχετης θεματολογίας.

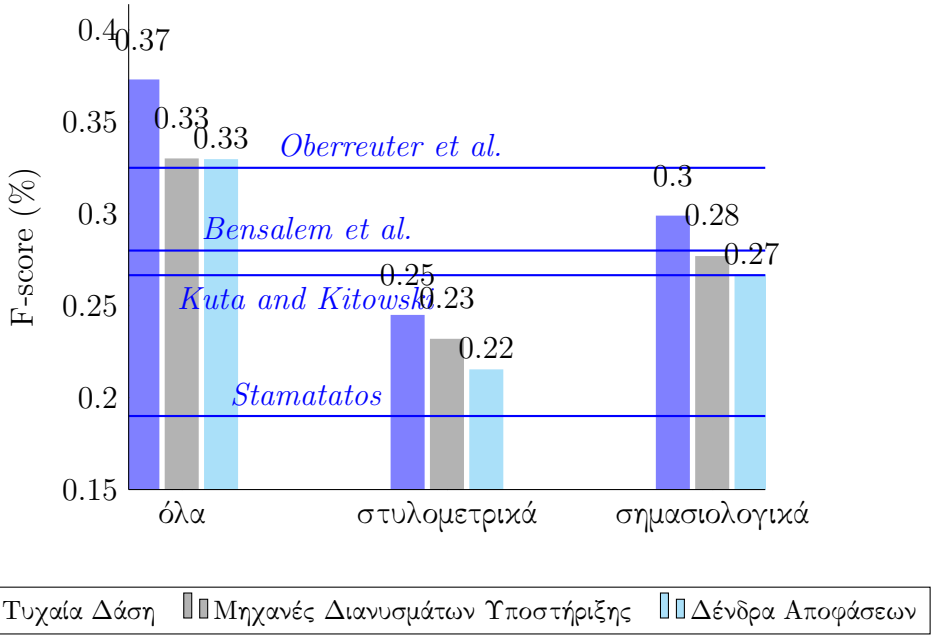
Και στα δύο σύνολα δεδομένων το σύστημά μας επιτυγχάνει τα καλύτερα αποτελέσματα. Συγκεκριμένα, στο σώμα κειμένων του 2009 επιτυγχάνουμε υψηλότερο F-score κατά τουλάχιστον 10% από τους καλύτερους συμμετέχοντες *(Oberreuter et al. 0.35, Stamatatos 0.25 F-score)*, από το σύστημα των Bensalem et al. *(F-score 0.33)* και από το σύστημα των Kuta και Kitowski *(F-score 0.3341)*.

Στο σύνολο κειμένων του 2011 το σύστημα μας επιτυγχάνει περίπου 5% υψηλότερο F-score στην περίπτωση του ταξινομητή τυχαίου δάσους, δίνοντας παράλληλα ελαφρώς καλύτερα αποτελέσματα για τους άλλους δύο ταξινομητές. Στο ίδιο σώμα κειμένων, το σύστημά μας φαίνεται ανώτερο από εκείνο των Stamatatos, Bensalem et al. και Kuta και Kitowski, επιτυγχάνοντας περίπου 18% υψηλότερη βαθμολογία στην πρώτη περίπτωση, 9% στη δεύτερη και 10% στην τρίτη.

Είναι προφανές ότι το σύνολο δεδομένων του 2011 ευνοεί περισσότερο τα σημασιολογικά χαρακτηριστικά. Στην περίπτωση του σώματος κειμένων του 2009 τα στυλομετρικά χαρακτηριστικά φαίνεται να είναι αρκετά ανταγωνιστικά, τόσο σε σύγκριση με τα σημασιολογικά χαρακτηριστικά, όσο και σε σχέση με τα άλλα συστήματα. Στις περισσότερες περιπτώσεις, ακόμη και με τα μισά χαρακτηριστικά, το σύστημά μας επιτυγχάνει καλύτερα αποτελέσματα. Αυτό



Σχήμα 9.4: F-score - PAN 2009



Σχήμα 9.5: F-score - PAN 2011

δεν ισχύει μόνο στην περίπτωση του συστήματος των Oberreuter et al., όταν δοκιμάζονται στο σύνολο δεδομένων του 2011. Αυτό συμβαίνει γιατί το συγκεκριμένο σύστημα ανίχνευσης λογοκλοπής εφαρμόζει ένα ισχυρό σημασιολογικό χαρακτηριστικό για την στυλιστική ανάλυση, αξιοποιώντας την αδυναμία του συνόλου δεδομένων.

Κεφάλαιο 10

Συμπεράσματα και μελλοντικές κατευθύνσεις έρευνας

10.1 Γενικά Συμπεράσματα

Στο πλαίσιο της παρούσας διατριβής, παρουσιάζονται τρεις μεθοδολογίες για την Ανίχνευση Κοινοτήτων και την Εγγενή Ανίχνευση Λογοκλοπής, οι οποίες κάνουν χρήση τεχνικών μηχανικής μάθησης, και βασίζονται στις ιδέες της ομοιότητας ή της ανομοιότητας.

Η πρώτη προτεινόμενη μεθοδολογία (Κεφάλαια 4, 5) αφορά τον εντοπισμό κοινοτήτων στο Twitter και την εξαγωγή ενδιαφερόντων ομάδων χρηστών. Η συγκεκριμένη προσέγγιση χρησιμοποιεί το περιεχόμενο που μοιράζονται οι χρήστες, τη δομή του γράφου και τις αλληλεπιδράσεις μεταξύ των χρηστών για τον ορισμό έξι μετρικών ομοιότητας.

Η διάδοση συνάφειας, ένας αλγόριθμος ομαδοποίησης που χρησιμοποιεί την ομοιότητα μεταξύ των σημείων δεδομένων και όχι τις απόλυτες θέσεις τους, χρησιμοποιήθηκε για την ομαδοποίηση των χρηστών σε κοινότητες. Ο αλγόριθμος LDA εκτελέστηκε προκειμένου να προσδιοριστεί η θεματολογία που συζητήθηκε από κάθε χρήστη, καθώς και το μείγμα θεμάτων κάθε ομάδας χρηστών. Το κριτήριο Calinski-Harabasz χρησιμοποιήθηκε για να βρεθούν οι τιμές των βαρών για κάθε μετρική ομοιότητας, γεγονός που έχει ως αποτέλεσμα την υψηλή ομοιότητα εντός των ομάδων και τη χαμηλη ομοιότητα ανάμεσα στις ομάδες.

Επιπλέον, προτάθηκε μια μέθοδος για την αφαίρεση των τετριμμένων θεμάτων και εξάχθηκαν αυτόματα οι επισημάνσεις για τα ενδιαφέροντα θέματα χρησιμοποιώντας την αγγλική Wikipedia. Τέλος, εισήχθη μια νέα μετρική που εκφράζει το ενδιαφέρον κάθε ομάδας.

Στη συνέχεια προτείνεται μια νέα, μη επιβλεπόμενη μεθοδολογία, βασισμένη

στους τυχαίους περίπατους, για την ενσωμάτωση των κόμβων ενός γράφου σε έναν διανυσματικό χώρο χαμηλών διαστάσεων (Κεφάλαια 6, 7). Σε αντίθεση με προηγούμενες εργασίες, η εν λόγω προσέγγιση λαμβάνει υπ΄ όψιν τόσο την πληροφορία σχετικά με τις ακμές ενός γράφου, όσο και τις ομοιότητες μεταξύ των κόμβων. Αναγκάζοντας τον τυχαίο περίπατο να μετακινηθεί σε κόμβους που είναι παρόμοιοι με τον τρέχοντα κόμβο, παράγονται πλουσιότερες αναπαραστάσεις των κόμβων, καθώς τα διανύσματα που προκύπτουν κωδικοποιούν επίσης τις ομοιότητες μεταξύ κάθε ζεύγους κόμβων.

Εισάγονται τρεις διαφορετικές διαδικασίες για την παραγωγή των τυχαίων περιπάτων, οι οποίες υλοποιούν (i) ομοιόμορφους τυχαίους περίπατους, (ii) μη ομοιόμορφους τυχαίους περίπατους σε γειτονικούς κόμβους, και (iii) μη ομοιόμορφους τυχαίους περίπατους σε οποιονδήποτε κόμβο. Επιπλέον, προτείνεται ένας τρόπος συνδυασμού αυτών των τριών διαδικασιών.

Η προτεινόμενη μεθοδολογία αξιολογείται σε έναν αριθμό τεχνητών και πραγματικών δικτύων. Τα αποτελέσματα δείχνουν ότι η μεθοδολογία λειτουργεί καλύτερα από τους state-of-the-art αλγορίθμους, όταν εφαρμόζεται σε δίκτυα που οι επισημάνσεις των κόμβων ευθυγραμμίζονται με τη δομή της κοινότητας των δικτύων. Σε όλες τις άλλες περιπτώσεις, τα αποτελέσματα ποικίλλουν ανάλογα με την επιλεγμένη στρατηγική. Ωστόσο, κάποιες στρατηγικές ξεπερνούν τους αλγόριθμους state-of-the-art, ακόμα και αυτούς που αποτελούν ημι-επιβλεπόμενες προσεγγίσεις.

Η στρατηγική με τις καλύτερες επιδόσεις συνολικά προκύπτει όταν οι πληροφορίες σχετικά με τις συνδέσεις και τις ομοιότητες κωδικοποιούνται σε διαφορετικά χαρακτηριστικά (διαφορετικές στήλες στον τελικό πίνακα ενσωμάτωσης), έτσι ώστε ο ταξινομητής που θα χρησιμοποιήσει αυτά τα χαρακτηριστικά να μπορεί να καθορίσει ποια είναι τα καλύτερα για το πρόβλημα υπό μελέτη.

Τέλος, εξετάζεται η απόδοση του συστήματος όταν χρησιμοποιούνται διαφορετικές μετρικές ομοιότητας. Τα αποτελέσματα δείχνουν ότι η επιλογή της μετρικής ομοιότητας δεν επηρεάζει σημαντικά τα αποτελέσματα, καθώς το σύστημα λειτουργεί εξίσου καλά ανεξάρτητα από την επιλογή της μετρικής.

Η τρίτη προτεινόμενη μεθοδολογία (Κεφάλαια 8, 9) αφορά την εγγενή ανίχνευση λογοκλοπής, δηλαδή την ανακάλυψη αποσπασμάτων σε ένα έγγραφο, τα οποία έχουν προκύψει από λογοκλοπή, μέσω του προσδιορισμού των στυλιστικών αλλαγών και των ασυμφωνιών μέσα στο ίδιο το έγγραφο, δεδομένου ότι δεν είναι διαθέσιμο το σώμα κειμένων αναφοράς. Η βασική ιδέα βασίζεται στην ανάλυση του προσωπικού τρόπου γραφής του αρχικού συγγραφέα και την επισήμανση των αποσπασμάτων που φαίνεται να διαφέρουν σημαντικά από αυτόν.

Η προσέγγιση προτείνει την εφαρμογή μιας ποικιλίας χαρακτηριστικών σε συνδυασμό με ένα κατάλληλο υποσύστημα μηχανικής μάθησης. Με αυτό τον τρόπο, είναι πιο πιθανή η ανακάλυψη μη εμφανών σχέσεων μεταξύ των χαρακτη-

ριστικών και η αναγνώριση στυλιστικών ανωμαλιών που τείνουν να παρουσιάζουν τα αποσπάσματα που έχουν προκύψει από λογοκλοπή. Τα νέα χαρακτηριστικά που προτείνονται είναι ιδιαίτερα ανταγωνιστικά σε σχέση με ήδη γνωστά χαρακτηριστικά, γεγονός που αποδεικνύεται από τα πειραματικά αποτελέσματα. Η σύγκριση με ήδη υπάρχοντα συστήματα ανίχνευσης εγγενούς λογοκλοπής δείχνει ότι το σύστημα επιτυγχάνει τα καλύτερα αποτελέσματα, ενώ σε ορισμένες περιπτώσεις καταφέρνει να ξεπερνά τις υπόλοιπες προσεγγίσεις ακόμα και όταν χρησιμοποιείται ένα υποσύνολο των προτεινόμενων χαρακτηριστικών.

Λόγω του σημαντικού ρόλου της μηχανικής μάθησης σε αυτή την προσέγγιση, το γεγονός ότι τα δεδομένα υπό εξέταση είναι μη ισορροπημένα αποτελεί κρίσιμη παράμετρο του προβλήματος. Για το λόγο αυτό, δοκιμάζονται διάφορες τεχνικές εξισορρόπησης των δεδομένων, καταλήγοντας στο συμπέρασμα ότι η εξισορρόπηση διαδραματίζει βασικό ρόλο για το αποτέλεσμα της ταξινόμησης, ενώ η τεχνική εξισορρόπησης πρέπει πάντα να επιλέγεται λαμβάνοντας υπ΄ όψιν το είδος του ταξινομητή.

Για την κατάτμηση των κειμένων χρησιμοποιείται η μέθοδος του μετακινούμενου παραθύρου. Εκτός από την κλασική μέθοδο, όπου το μήκος του παραθύρου και το βήμα είναι σταθερά, δοκιμάζονται διαφορετικές τιμές για τις δύο αυτές παραμέτρους, οι οποίες εξαρτώνται από το μέγεθος του ύποπτου εγγράφου. Ενώ η ιδέα της προσαρμογής του μεγέθους παραθύρου στο μήκος του εγγράφου φαίνεται, υποσχόμενη, τα πειραματικά αποτελέσματα αποδεικνύουν ότι η κλασική μέθοδος δίνει καλύτερα αποτελέσματα.

10.2 Μελλοντικές Επεκτάσεις

Από την παρούσα διατριβή προκύπτουν ορισμένα ζητήματα τα οποία θα μπορούσαν να αποτελέσουν αντικείμενο μελέτης επόμενων εργασιών.

Ένα επόμενο πιθανό βήμα θα ήταν η χρήση των διανυσματικών παραστάσεων κόμβων στα κοινωνικά δίκτυα, ή γενικότερα σε δίκτυα τα οποία περιλαμβάνουν τόσο συνδέσεις ανάμεσα σε αντικείμενα, όσο και άλλου είδους αλληλεπιδράσεις. Όπως εξηγήθηκε στο κεφάλαιο 4, τα κοινωνικά δίκτυα, πέρα από τις ακμές, δηλαδή τις κοινωνικές συνδέσεις ανάμεσα στους χρήστες/κόμβους, έχουν και άλλα χαρακτηριστικά τα οποία μπορούν να δώσουν πληροφορία σε σχέση με την ομοιότητα των κόμβων. Όλες αυτές οι ομοιότητες μπορούν να χρησιμοποιηθούν κατά την εξερεύνηση των δικτύων, και κατ΄ επέκταση να κωδικοποιηθούν με τη μορφή διανυσμάτων.

Μια άλλη πιθανή επέκταση αφορά τη δεύτερη μεθοδολογία, η οποία περιγράφεται στο κεφάλαιο 6. Όπως εξηγήθηκε στην περιγραφή των συνόλων δεδομένων (ενότητα 7.1), κάθε δίκτυο παρουσιάζει μείγμα ομοφιλίας και ισοδυναμίας ρόλων. Μέχρι στιγμής, η συγκεκριμένη προσέγγιση είναι στραμμένη προς την

ανίχνευση κοινοτήτων, επομένως δεν έχει μελετηθεί η ισοδυναμία ρόλων. Ένας τρόπος για να επεκταθεί η μεθοδολογία, ώστε να κωδικοποιεί περισσότερες από τις ιδιότητες του δικτύου, θα ήταν να χρησιμοποιηθεί κάποιο κριτήριο που να εκφράζει την ισοδυναμία ρόλων μεταξύ των κόμβων, συνδυαστικά με τις στρατηγικές ενσωμάτωσης που έχουν ήδη προταθεί.

Bibliography

[1] Project gutenberg. http://www.gutenberg.org.

[2] Aizerman, M. A. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control 25* (1964), 821–837.

[3] Alsallal, M., Iqbal, R., Amin, S., and James, A. Intrinsic plagiarism detection using latent semantic indexing and stylometry. In *2013 Sixth International Conference on Developments in eSystems Engineering* (Dec 2013), pp. 145–150.

[4] Alzahrani, S., Salim, N., and Abraham, A. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42* (2012), 133–149.

[5] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl. 6*, 1 (June 2004), 20–29.

[6] Belkin, M., and Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Cambridge, MA, USA, 2001), NIPS'01, MIT Press, pp. 585–591.

[7] Bensalem, I., Rosso, P., and Chikhi, S. Intrinsic plagiarism detection using n-gram classes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (2014), pp. 1459–1464.

[8] Blei, D. M. Probabilistic topic models. *Commun. ACM 55*, 4 (Apr. 2012), 77–84.

[9] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[10] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008*, 10 (oct 2008), P10008.

[11] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (New York, NY, USA, 1992), COLT '92, ACM, pp. 144–152.

[12] Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *CoRR abs/1106.1813* (2011).

[13] Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct 2001), 5–32.

[14] Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. Classification and regression trees chapman & hall. *New York* (1984).

[15] Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bähler, J., Wood, V., et al. The biogrid interaction database: 2008 update. *Nucleic acids research 36*, suppl_1 (2007), D637–D640.

[16] Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*, 2 (Jun 1998), 121–167.

[17] Caliński, T., and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation 3*, 1 (1974), 1–27.

[18] Cavallari, S., Zheng, V. W., Cai, H., Chang, K. C.-C., and Cambria, E. Learning community embedding with community detection and node embedding on graphs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), ACM, pp. 377–386.

[19] Chen, H., Perozzi, B., Hu, Y., and Skiena, S. HARP: hierarchical representation learning for networks. *CoRR abs/1706.07845* (2017).

[20] Cheng, N., Chandramouli, R., and Subbalakshmi, K. P. Author gender identification from text. *Digit. Investig. 8*, 1 (July 2011), 78–88.

[21] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[22] Courant, R., and Hilbert, D. Methods of mathematical physics, new york, interscience, 1953. *Mathematical Reviews (MathSciNet): MR16: 426a Zentralblatt MATH 53*.

[23] Curran, D. An evolutionary neural network approach to intrinsic plagiarism detection. In *Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science* (Berlin, Heidelberg, 2010), AICS'09, Springer-Verlag, pp. 33–40.

[24] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 41*, 6 (1990), 391–407.

[25] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018).

[26] Ding, Y. Community detection: Topological vs. topical. *Journal of Informetrics 5*, 4 (2011), 498–514.

[27] Dubay, W. H. The principles of readability. *Costa Mesa, CA: Impact Information* (2004).

[28] Eissen, S. M. z., and Stein, B. Intrinsic plagiarism detection. In *Advances in Information Retrieval* (Berlin, Heidelberg, 2006), M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, Eds., Springer Berlin Heidelberg, pp. 565–569.

[29] Figueiredo, D. R., Ribeiro, L. F. R., and Saverese, P. H. P. struc2vec: Learning node representations from structural identity. *CoRR abs/1704.03165* (2017).

[30] Firth, J. R. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).

[31] Fortunato, S. Community detection in graphs. *Physics reports 486*, 3-5 (2010), 75–174.

[32] Frey, B. J., and Dueck, D. Clustering by passing messages between data points. *Science 315* (2007), 972–976.

[33] Girvan, M., and Newman, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences 99*, 12 (2002), 7821–7826.

[34] Goel, A., Sharma, A., Wang, D., and Yin, Z. Discovering similar users on twitter. In *11th Workshop on Mining and Learning with Graphs* (2013).

[35] Goyal, P., and Ferrara, E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems 151* (2018), 78–94.

[36] Πολυδούρη, Α. Εγγενής ανίχνευση λογοκλοπής με ευφυείς τεχνικές, Διπλωματική Εργασία, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο, 2016.

[37] Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National academy of Sciences 101*, suppl 1 (2004), 5228–5235.

[38] Grover, A., and Leskovec, J. node2vec: Scalable feature learning for networks. *CoRR abs/1607.00653* (2016).

[39] Gunasekaran, K., Muralikumar, J., Sudarshan, S., Srinivasan, B., and Malliaros, F. Netglove: Learning node representations for community detection. In *6th International Conference on Complex Networks and Their Applications* (2017).

[40] Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems* (2017), pp. 1024–1034.

[41] Hamilton, W. L., and Tang, J. Graph representation learning. AAAI-19 Tutorial Forum, 2019.

[42] Hamilton, W. L., Ying, R., and Leskovec, J. Representation learning on graphs: Methods and applications. *CoRR abs/1709.05584* (2017).

[43] Harris, Z. Distributional structure. word, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), readings in the philosophy of language, 1954.

[44] Ho, T. K. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1* (Washington, DC, USA, 1995), ICDAR '95, IEEE Computer Society, pp. 278–.

[45] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Computation 9*, 8 (1997), 1735–1780.

[46] Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1999), SIGIR '99, ACM, pp. 50–57.

[47] Holmes, D. I. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing 13*, 3 (1998), 111–117.

[48] Hua, X., Li, S., Li, P., and Zhu, Q. Research on intrinsic plagiarism detection resolution: A supervised learning approach. In *Chinese Lexical Semantics* (Berlin, Heidelberg, 2013), D. Ji and G. Xiao, Eds., Springer Berlin Heidelberg, pp. 58–63.

[49] Hyafil, L., and Rivest, R. L. Constructing optimal binary decision trees is np-complete. *Inf. Process. Lett. 5* (1976), 15–17.

[50] Jolliffe, I. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[51] Kestemont, M., Luyckx, K., and Daelemans, W. Intrinsic plagiarism detection using character trigram distance scores. In *Notebook for PAN at CLEF 2011* (2011).

[52] Kim, J., and Lee, J.-G. Community detection in multi-layer graphs: A survey. *SIGMOD Rec. 44*, 3 (Dec. 2015), 37–48.

[53] Koppel, M., and Schler, J. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-first International*

Conference on Machine Learning (New York, NY, USA, 2004), ICML '04, ACM, pp. 62–.

[54] Kuta, M., and Kitowski, J. Optimisation of character n-gram profiles method for intrinsic plagiarism detection. In *Artificial Intelligence and Soft Computing* (Cham, 2014), L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Springer International Publishing, pp. 500–511.

[55] Kuznetsov, M., Motrenko, A., Kuznetsova, R., and Strijov, V. Methods for intrinsic plagiarism detection and author diarization. In *CLEF* (2016).

[56] Lancichinetti, A., Fortunato, S., and Radicchi, F. Benchmark graphs for testing community detection algorithms. *Physical review E 78*, 4 (2008), 046110.

[57] Li, D., Ding, Y., Shuai, X., Bollen, J., Tang, J., Chen, S., Zhu, J., and Rocha, G. Adding community and dynamic to topic models. *Journal of Informetrics 6*, 2 (2012), 237–253.

[58] Li, H. J. The comparison of significance of fuzzy community partition across optimization methods. *Journal of Intelligent & Fuzzy Systems 29*, 6 (2015), 2707–2715.

[59] Lim, K. H., and Datta, A. Tweets beget propinquity: Detecting highly interactive communities on twitter using tweeting links. In *IEEE/WIC/ACM International Conference on Web Intelligence* (2012), vol. 1, pp. 214–221.

[60] Lior, R., and Maimon, O. Clustering methods. *Data mining and knowledge discovery handbook. Springer US* (2005), 321–352.

[61] Liu, W., Deng, Z. H., Cao, L., Xu, X., Liu, H., and Gong, X. Mining top k spread sources for a specific topic and a given node. *IEEE Transactions on Cybernetics 45*, 11 (2015), 2472–2483.

[62] Mahoney, M. Large text compression benchmark, 2011. www.mattmahoney.net/dc/textdata.

[63] Malliaros, F. D., and Vazirgiannis, M. Clustering and community detection in directed networks: A survey. *CoRR abs/1308.0971* (2013).

[64] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[65] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[66] Morin, F., and Bengio, Y. Hierarchical probabilistic neural network language model. In *Aistats* (2005), vol. 5, Citeseer, pp. 246–252.

[67] Oberreuter, G., L'Huillier, G., Ríos, S. A., and Velásquez, J. D. Approaches for intrinsic and external plagiarism detection - notebook for pan at clef 2011. In *CLEF* (2011).

[68] Oberreuter, G., and Velásquez, J. D. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications 40*, 9 (2013), 3756 – 3763.

[69] Ou, M., Cui, P., Pei, J., Zhang, Z., and Zhu, W. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 1105–1114.

[70] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. Community detection in social media. *Data Mining and Knowledge Discovery 24*, 3 (May 2012), 515–554.

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

[72] Perez-Tellez, F., Cardiff, J., Rosso, P., and Pinto, D. Weblog and short text feature extraction and impact on categorisation. *Journal of Intelligent & Fuzzy Systems 27*, 5 (2014), 2529–2544.

[73] Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. *CoRR abs/1403.6652* (2014).

[74] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *CoRR abs/1802.05365* (2018).

[75] Polydouri, A., Siolas, G., and Stafylopatis, A. Intrinsic plagiarism detection with feature-rich imbalanced dataset learning. In *Engineering Applications of Neural Networks* (Cham, 2017), G. Boracchi, L. Iliadis, C. Jayne, and A. Likas, Eds., Springer International Publishing, pp. 99–110.

[76] Potthast, M., Eiselt, A., Barrón-cedeño, A., Stein, B., and Rosso, P. Overview of the 3rd international competition on plagiarism detection. In *In Working Notes Papers of the CLEF 2011 Evaluation* (2011).

[77] Potthast, M., Stein, B., Eiselt, A., universität Weimar, B., Barrón-cedeño, A., and Rosso, P. P.: Overview of the 1st international competition on plagiarism detection. In *In: SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org* (2009), pp. 1–9.

[78] Pyo, S., Kim, E., and Kim, M. Lda-based unified topic modeling for similar tv user grouping and tv program recommendation. *IEEE Transactions on Cybernetics 45*, 8 (2015), 1476–1490.

[79] Quinlan, J. R. Induction of decision trees. *Mach. Learn. 1*, 1 (Mar. 1986), 81–106.

[80] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[81] Raghavan, U. N., Albert, R., and Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E 76*, 3 (Sep 2007).

[82] Ranatunga, R. V. S. P. K., Atukorale, A. S., and Hewagamage, K. P. Intrinsic plagiarism detection with kohonen self organizing maps. In *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)* (Sept 2011), pp. 125–125.

[83] Rokach, L., and Maimon, O. *Data Mining With Decision Trees: Theory and Applications*, 2nd ed. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2014.

[84] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* (1958), 65–386.

[85] Roshchina, A., Cardiff, J., and Rosso, P. Twin: Personality-based intelligent recommender system. *Journal of Intelligent & Fuzzy Systems 28*, 5 (2015), 2059–2071.

[86] Rosso, P., Pardo, F. M. R., Potthast, M., Stamatatos, E., Tschuggnall, M., and Stein, B. Overview of pan'16 - new challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In *CLEF* (2016).

[87] Roweis, S. T., and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science 290*, 5500 (2000), 2323–2326.

[88] Ruan, Y., Fuhry, D., and Parthasarathy, S. Efficient community detection in large networks using content and links. In *22nd International Conference on World Wide Web* (2013), pp. 1089–1098.

[89] Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. Learning representations by back-propagating errors. *Cognitive modeling 5*, 3 (1988), 1.

[90] Salton, G., and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[91] Salton, G., Wong, A., and Yang, C.-S. A vector space model for automatic indexing. *Communications of the ACM 18*, 11 (1975), 613–620.

[92] Seaward, L., and Matwin, S. Intrinsic plagiarism detection using complexity analysis. In *Proc. SEPLN* (2009), pp. 56–61.

[93] Sittar, A., Iqbal, H. R., and Nawab, R. M. A. Author diarization using cluster-distance approach. In *CLEF (Working Notes)* (2016), pp. 1000–1007.

[94] Stamatatos, E. Intrinsic plagiarism detection using character n-gram profiles. In *SEPLN 2009 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09)* (2009), S. B., R. P., S. E., K. M., and A. E., Eds., pp. 38–46.

[95] Stamatatos, E. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol. 60*, 3 (Mar. 2009), 538–556.

[96] Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., and Stein, B. Overview of the author identification task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers* (Toulouse, France, 2015/09/10 2015), CEUR, CEUR.

[97] Steenfatt, N., Nikolentzos, G., Vazirgiannis, M., and Zhao, Q. Learning structural node representations on directed graphs. In *Complex Networks and Their Applications VII - Volume 2 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018* (2018), pp. 132–144.

[98] Stein, B., Lipka, N., and Prettenhofer, P. Intrinsic plagiarism analysis. *Lang. Resour. Eval. 45*, 1 (Mar. 2011), 63–82.

[99] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web* (2015), International World Wide Web Conferences Steering Committee, pp. 1067–1077.

[100] Tang, L., and Liu, H. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 817–826.

[101] Tang, Y., Zhang, Y., Chawla, N. V., and Krasser, S. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39* (2009), 281–288.

[102] Tenenbaum, J. B., De Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science 290*, 5500 (2000), 2319–2323.

[103] Tschuggnall, M., and Specht, G. Plag-inn: Intrinsic plagiarism detection using grammar trees. In *Natural Language Processing and Information Systems* (Berlin, Heidelberg, 2012), G. Bouma, A. Ittoo, E. Métais, and H. Wortmann, Eds., Springer Berlin Heidelberg, pp. 284–289.

[104] Tschuggnall, M., and Specht, G. Using grammar-profiles to intrinsically expose plagiarism in text documents. In *Natural Language Processing and Information Systems* (Berlin, Heidelberg, 2013),

E. Métais, F. Meziane, M. Saraee, V. Sugumaran, and S. Vadera, Eds., Springer Berlin Heidelberg, pp. 297–302.

[105] Vapnik, V. The nature of statistical learning theory· 6·[mj new york. *Springer-Verlag 1* (1995), 995.

[106] Vapnik, V. N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[107] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR abs/1706.03762* (2017).

[108] Vathi, E., Siolas, G., and Stafylopatis, A. Mining and categorizing interesting topics in twitter communities. *Journal of Intelligent & Fuzzy Systems 32*, 2 (2017), 1265–1275.

[109] von Luxburg, U. A tutorial on spectral clustering. *CoRR abs/0711.0189* (2007).

[110] Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., and Yang, S. Community preserving network embedding. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

[111] Weng, J., Lim, E. P., Jiang, J., and He, Q. TwitterRank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (2010), pp. 261–270.

[112] Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics SMC-2*, 3 (July 1972), 408–421.

[113] Xie, J., Kelley, S., and Szymanski, B. K. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys 45*, 4 (2013), 1–35.

[114] Yang, J., McAuley, J. J., and Leskovec, J. Community detection in networks with node attributes. In *IEEE 13th International Conference on Data Mining* (2013), pp. 1151–1156.

[115] Zhang, Y., Wu, Y., and Yang, Q. Community discovery in twitter based on user interests. *Journal of Computational Information Systems 8*, 3 (2012), 991–1000.

[116] Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., and Fan, J. Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems 26* (2012), 164–173.

[117] Zhu, X., and Ghahramani, Z. Learning from labeled and unlabeled data with label propagation.

Γλωσσάριο τεχνικών όρων

Accuracy	Ορθότητα
Activation function	Συνάρτηση ενεργοποίησης
Affinity Propagation	Διάδοση Συνάφειας
Artificial Intelligence	Τεχνητή Νοημοσύνη
Artificial Neural Networks	Τεχνητά Νευρωνικά Δίκτυα
Author Diarization	Διαφοροποίηση Συγγραφέα
Author Identification	Αναγνώριση Συγγραφέα
Author Verification	Επαλήθευση Συγγραφέα
Author Attribution	Απόδοση Συγγραφέα
Backpropagation	Οπισθοδιάδοση
Balancing technique	Τεχνική εξισορρόπησης
Between-group sum of squares	Άθροισμα τετραγώνων μεταξύ των ομάδων
Bias	Πόλωση
Biased random walk	Μεροληπτικός τυχαίος περίπατος
Bootstrap aggregating (bagging)	Πολλαπλή δειγματοθέτηση
Classification	Ταξινόμηση
Classifier	Ταξινομητής
Clique	Κλίκα
Cluster	Ομάδα/Συστάδα
Clustering	Ομαδοποίηση/Συσταδοποίηση
Collection Topic Distribution	Κατανομή Θεμάτων Συλλογής
Community Detection	Ανίχνευση κοινοτήτων
Complete graph	Πλήρης γράφος
Complete mutuality	Πλήρης αμοιβαιότητα
Composite Ranking Index	Σύνθετος Δείκτης Κατάταξης
Compression rate	Ποσοστό συμπίεσης
Concatenation	Συνένωση
Correlated features	Συσχετισμένα χαρακτηριστικά
Data Mining	Εξόρυξη Γνώσης από Δεδομένα
Decision Trees	Δένδρα Αποφάσεων
Deep Learning	Βαθιά Μάθηση

Degree	Βαθμός
Dimensionality reduction	Μείωση διαστάσεων
Directed graph	Κατευθυνόμενος γράφος
Distributional hypothesis	Υπόθεση κατανομής
Edge betweenness centrality	Κεντρικότητα της ενδιαμεσότητας ακμών
Edited Nearest Neighbors	Επεξεργασμένοι Πλησιέστεροι Γείτονες
Embedding weights	Βάρη ενσωμάτωσης
Ensemble learning	Μάθηση συνόλου/Συνδυαστική μάθηση
Exemplar	Υπόδειγμα/Πρότυπο
External degree	Εξωτερικός βαθμός
False negatives	Ψευδή αρνητικά
False positives	Ψευδή θετικά
Feature learning	Μάθηση χαρακτηριστικών
Feature vector	Διάνυσμα χαρακτηριστικών
Flesch reading-ease test	Έλεγχος αναγνωσιμότητας του Flesch
Follower	Ακόλουθος
Following relationship	Σχέση ακολούθησης
Forward pass	Ευθύ πέρασμα
Friend	Φίλος
Gibbs sampling	Δειγματοληψία Gibbs
Gradient descent	Κάθοδος κλίσης
Grammar tree	Δέντρο γραμματικής
Graph embedding	Ενσωμάτωση/Ένθεση γράφου
Graph/network clustering	Ομαδοποίηση γράφου/δικτύου
Grid-search	Αναζήτηση πλέγματος
Information gain	Κέρδος πληροφορίας
Information Retrieval	Ανάκτηση Πληροφορίας
Inter-cluster density	Δια-συσταδική πυκνότητα
Internal degree	Εσωτερικός βαθμός
Intra-cluster density	Ενδο-συσταδική πυκνότητα
Intrinsic Plagiarism Detection	Εγγενής Ανίχνευση Λογοκλοπής
Kernel function	Συνάρτηση πυρήνα
Kolmogorov Complexity	Πολυπλοκότητα Kolmogorov
Label	Επισήμανση
Lagrangian formulation	Λαγκρανζιανή διατύπωση
Latent Dirichlet allocation	Λανθάνουσα Ανάθεση Dirichlet
Latent Semantic Indexing	Λανθάνουσα Σημασιολογική Δεικτοδότηση
Lemmatization	Λημματοποίηση
Local Topic Distribution	Τοπική Κατανομή Θεμάτων
Local Topic Interestingness	Ενδιαφέρον Τοπικής Θεματολογίας

Local Twitter Community Interestingness	Ενδιαφέρον Τοπικών Κοινοτήτων του Twitter
Logistic regression	Λογιστική παλινδρόμηση
Logistic sigmoid	Λογιστική σιγμοειδής
Machine Learning	Μηχανική Μάθηση
Margin	Περιθώριο
Marginal distribution	Περιθώρια κατανομή
Maximum marginal hyperplane	Μέγιστο περιθώριο υπερεπίπεδο
Mean Cluster Percentage	Μέσο Ποσοστό ανά Ομάδα
Mean Word Frequency	Μέση Συχνότητα Λέξεων
Mention	Αναφορά
Misclassification error	Λάθος ταξινόμησης
Mixing parameter	Παράμετρος μίξης
Multi-class classification	Ταξινόμηση πολλαπλών κλάσεων
Multi-label classification	Ταξινόμηση πολλαπλών ετικετών
Multilayer Perceptron	Πολυστρωματικά Perceptron
Multinomial distribution	Πολυωνυμική κατανομή
Named entity recognition	Αναγνώριση ονομάτων οντοτήτων
Natural Language Processing	Επεξεργασία Φυσικής Γλώσσας
Negative sampling	Δειγματοληψία αρνητικών
Node embedding	Διανυσματική παράσταση κόμβου
Non-probabilistic	Μη-πιθανοτικός
Non-uniform random walk	Μη ομοιόμορφος τυχαίος περίπατος
One-vs-rest	Ένα-έναντι-υπολοίπων
Order of network proximity	Τάξη εγγύτητας δικτύου
Overfitting	Υπερπροσαρμογή
Oversampling	Υπερδειγματοληψία
Part-of-speech tagging	Επισήμανση μερών του λόγου
Plagiarism	Λογοκλοπή
Precision	Ακρίβεια
Probabilistic Latent Semantic Indexing	Πιθανοτική Λανθάνουσα Σημασιολογική Δεικτοδότηση
Pruning	Κλάδεμα
Random Forest	Τυχαίο Δάσος
Reachability	Προσβασιμότητα
Recall	Ανάκληση
Regression	Παλινδρόμηση
Reinforcement learning	Ενισχυτική μάθηση
Reply	Απάντηση
Reverse pass	Ανάστροφο πέρασμα
Run length encoding	Κωδικοποίηση μήκους διαδρομής

Self-Organizing Map	Αυτο-οργανούμενος Χάρτης
Semi-supervised learning	Ημι-επιβλεπόμενη μάθηση
Sentiment analysis	Ανάλυση συναισθήματος
Similarity	Ομοιότητα
Single-class classification	Ταξινόμηση μιας κλάσης
Singular Value Decomposition	Αποσύνθεση Ιδιαζουσών Τιμών
Slack variables	Μεταβλητές χαλαρότητας
Stemming	Αποκοπή καταλήξεων/Αναγωγή στο θέμα
Stopwords	Συχνές λέξεις
Stylometry	Στυλομετρία
Supervised learning	Επιβλεπόμενη μάθηση
Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
Support vectors	Διανύσματα υποστήριξης
Test set	Σύνολο ελέγχου
Text Mining	Εξόρυξη Γνώσης από Κείμενο
Text preprocessing	Προεπεξεργασία κειμένου
Tokenization	Ανίχνευση λεκτικών μονάδων
Topic Modelling	Μοντελοποίηση Θεμάτων/Θεματική Μοντελοποίηση
Topic-based approaches	Προσεγγίσεις που βασίζονται στη θεματολογία
Topology-based approaches	Προσεγγίσεις που βασίζονται στην τοπολογία
Training set	Σύνολο εκπαίδευσης
True negatives	Αληθή αρνητικά
True positives	Αληθή θετικά
Undersampling	Υποδειγματοληψία
Undirected graph	Μη κατευθυνόμενος γράφος
Uniform random walk	Ομοιόμορφος τυχαίος περίπατος
Unsupervised learning	Μη επιβλεπόμενη μάθηση
User Topic Distribution	Κατανομή Θεμάτων Χρήστη
Variance ratio criterion	Κριτήριο αναλογίας διακύμανσης
Variance reduction	Μείωση της διακύμανσης
Vector space model	Μοντέλο διανυσματικού χώρου
Weighted graph	Σταθμισμένος γράφος
Within-group sum of squares	Άθροισμα τετραγώνων εντός των ομάδων
Word embedding	Διανυσματική παράσταση λέξης
Word frequency class	Κλάση συχνότητας λέξης

Κατάλογος δημοσιεύσεων

- Δημοσιεύσεις σε διεθνή περιοδικά με κρίση

    1. Eleni Vathi, Thanos Tagaris, Georgios Alexandridis, Andreas Stafylopatis: **Unsupervised Graph Embedding based on Node Similarity** (εκκρεμεί)
    2. Andrianna Polydouri, Eleni Vathi, Georgios Siolas, Andreas Stafylopatis : **An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection**, Evolving Systems (2018)
    3. Eleni Vathi, Georgios Siolas, Andreas Stafylopatis : **Mining and categorizing interesting topics in Twitter communities**, Journal of Intelligent and Fuzzy Systems 32(2): 1265-1275 (2017)

- Δημοσιεύσεις σε διεθνή συνέδρια με κρίση

    1. Eleni Vathi, Georgios Siolas, Andreas Stafylopatis : **Mining Interesting Topics in Twitter Communities**, 7th International Conference on Computational Collective Intelligence Technologies and Applications (2015)

- Δημοσιεύσεις εκτός διατριβής

    1. Vasileios Chasanis, Costas Voglis, Antonis Ioannidis, Aris Lanaridis, Eleni Vathi, Georgios Siolas, Aristidis Likas, Andreas Stafylopatis : **Videosum: a video storing, processing and summarization platform**, Conference for Visual Media Production (2014)