ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανίχνευση των τάσεων της αγοράς από διαδικτυακές πηγές και κοινωνικά μέσα για ένα συγκεκριμένο κλάδο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νεφέλη Παναγιωτοπούλου

Επιβλέπων : Δημήτριος Ασκούνης

Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανίχνευση των τάσεων της αγοράς από διαδικτυακές πηγές και κοινωνικά μέσα για ένα συγκεκριμένο κλάδο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νεφέλη Παναγιωτοπούλου

Επιβλέπων : Δημήτριος Ασκούνης

Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ……

....

Αθήνα, Ιούλιος 2017

.....

Νεφέλη Παναγιωτοπούλου

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

Copyright © Νεφέλη Παναγιωτοπούλου, 2017.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε υπό την επίβλεψη του Καθηγητή Δημήτρη Ασκούνη στο Εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου.

Το αντικείμενο της εργασίας επικεντρώνεται στην μελέτη των μεθόδων ανίχνευσης τάσεων στο διαδίκτυο και τα κοινωνικά μέσα, ενώ γίνεται εξειδίκευση και πρόταση μεθοδολογίας για τις τάσεις της αγοράς για ένα συγκεκριμένο κλάδο. Ο κλάδος που επιλέχθηκε είναι αυτός της βιομηχανίας κατασκευής συμβατικών παραδοσιακών παιχνιδιών. Η μεθοδολογία που προτείνεται μετά την υλοποίησή της μπορεί να αποτελέσει ένα εργαλείο επιχειρηματικής ευφυΐας και να αξιοποιηθεί ως τέτοιο με την ανάλογη παραμετροποίηση.

Θα ήθελα να ευχαριστήσω ιδιαίτερα τους επιβλέποντες της εργασίας μου Αριάδνη Μιχαλίτση-Ψαρρού και Δημήτρη Πανόπουλο για την δυνατότητα που μου έδωσαν να ασχοληθώ με το συγκεκριμένο θέμα, καθώς και για την συνεχή υποστήριξη και καθοδήγησή τους καθ'όλη τη διάρκεια εκπόνησης της διπλωματικής εργασίας.

Νεφέλη Παναγιωτοπούλου

Ιούλιος 2017

Περίληψη

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, μελετάται η τεχνολογική πρόοδος στον χώρο της ανίχνευσης τάσεων στα κοινωνικά μέσα και τον ιστό. Σαν αποτέλεσμα, προτείνεται μια μεθοδολογία για την ανίχνευση των τάσεων της αγοράς για ένα συγκεκριμένο κλάδο – εν προκειμένω, αυτόν της βιομηχανιίας κατασκευής παιχνιδιών. Μετά την πραγματοποίηση μελέτης της σχετικής βιβλιογραφίας, επιλέγεται ο κατάλληλος συνδυασμός σύγχρονων τεχνικών για τον εντοπισμό των τάσεων της αγοράς στο συγκεκριμένο κλάδο μέσα από τα κοινωνικά δίκτυα και άλλες διαδικτυακές πηγές. Από το συνδυασμό των τεχνικών θα μπορεί να σχεδιαστεί ένα ολοκληρωμένο σύστημα ανίχνευσης τάσεων, το οποίο μπορεί να χρησιμεύσει σαν ένα εργαλείο επιχειρηματικής ευφυΐας στα χέρια μια εταιρίας αυτού του τομέα για την αναγνώριση πιθανών «κενών» ή ευκαιριών της εξεταζόμενης αγοράς, προς τα οποία μπορεί να προσανατολιστεί.

Η έρευνα αγοράς και προτιμήσεων καταναλωτή εξετάζεται αριχικά σε θεωρητικό επίπεδο, από τη σκοπιά του κλάδου του μάρκετινγκ. Επίσης, δίνεται μια γενική εικόνα για τις σύγχρονες κατευθύνσεις στον τομέα της έρευνας αγοράς. Στη συνέχεια, μελετάται η ανίχνευση των τάσεων της αγοράς από διαδικτυακές πηγές από τεχνική σκοπιά και γίνεται αναφορά στα βασικά σημεία του γενικότερου τεχνικού υπόβαθρου που απαιτείται για την κατανόηση του ερευνητικού αντικειμένου και τη συμβολή σε αυτό (Επιχειρηματική Ευφυΐα, Εξόρυξη Δεδομένων, Μεγάλα Δεδομένα). Τέλος, γίνεται βιβλιογραφική επισκόπηση των μεθόδων που αφορούν το θέμα μας, δηλαδή την κατασκευή συστημάτων Επιχειρηματικής Ευφυΐας βασισμένων σε μεθόδους Εξόρυξης Δεδομένων, ειδικά όμως για την ανίχνευση των τάσεων της αγοράς. Για τους σκοπούς της έρευνας, συγκεντρώθηκαν πάνω από 50 μεθοδολογίες, 22 εκ των οποίων τα χαρακτηριστικά αποτυπώθηκαν στον σχετικό πίνακα στο παράρτημα της εργασίας. Το δείγμα αυτό θεωρείται αντιπροσωπευτικό του συνόλου.

Τα αποτελέσματα της έρευνας συγκεντρώνονται σε δύο άξονες. Πρώτον, γίνεται μια κατηγοριοποίηση των μεθοδολογιών και των βημάτων που ακολουθούν, καθώς και των τεχνικών που χρησιμοποιούνται. Δεύτερον, στο τελευταίο μέρος της εργασίας εξετάζεται κατά πόσο είναι χρήσιμη η εξόρυξη δεδομένων από τα κοινωνικά δίκτυα και το διαδίκτυο για τον συγκεκριμένο κλάδο, και προτείνεται ένα ολοκληρωμένο σύστημα ανίχνευσης των τάσεων της αγοράς από τις πηγές αυτές για το συγκεκριμένο κλάδο.

Λέξεις-κλειδιά: τάσεις, τάσεις αγοράς, ανίχνευση τάσεων, εξόρυξη δεδομένων, επιχειρηματική ευφυΐα, κατασκευή παιχνιδιών, βιομηχανία παιχνιδιών

Abstract:

For the purposes of the current diploma thesis, we have studied the progress in the field of trend detection on social media and other internet sources. We also propose a methodology for detecting market trends for a specific industry – the toy manufacturing industry. After having studied the related papers, we propose an appropriate combination of modern techniques that can perform trend detection on the social media and other web sources. The combination of techniques and algorithms can lead to the implementation of a market trend detection system. That system if implemented would be a valuable business intelligence tool in the hands of a company. It would help in the decision making process by identifying opportunities and dangers in the external environment of the company.

In the study, the market and consumer research is initially examined at an abstract level and in marketing terms. We also mention the directions of modern market and consumer research. Market trend detection from the social media and other internet sources is studied from a technical scope and we point out the basic technical background that is necessary for the understanding of our research field (Business Intelligence, Data Mining, Big Data). We studied extensively the existing set of methodologies that approach diferent forms of trend detection, that use Data Mining and other processes. For the purposes of that study we gathered more than 50 journal articles that propose a methodology, from which 22 were used to map and categorize the common stages and techniques used in the methodologies. We consider this sample to be representative of the whole.

The conclusions of the research are the following: Firsly, the categorization and analysis of methodologies for trend detection, the techniques used and the steps followed. Secondly, in the last part of the study we have examined how useful is the process for the toy manufacturing industry and proposed a market trend detection system for the industry, based on the previous work.

Keywords: trends, market trends, trend detection, data mining, business intelligence, toy manufacturing, toy industry

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΣΧΗΜΑΤΩΝ

1. Εισαγωγή

1.1. Στόχος της διπλωματικής εργασίας

Μια επιχείρηση για να επιβιώσει στο σύγχρονο ανταγωνιστικό περιβάλλον οφείλει να αφουγκράζεται διαρκώς τις ανάγκες των πελατών της, να εντοπίζει εγκαίρως τις νέες τάσεις της αγοράς και να προσαρμόζεται γρήγορα σε αυτές. Για να το κάνει αυτό πρέπει να μπορεί να αξιοποιεί κάθε είδους πληροφορία - από τα δεδομένα που η ίδια διαθέτει αλλά κυρίως από το εξωτερικό της περιβάλλον. Με την έρευνα που έχει πραγματοποιηθεί τα τελευταία χρόνια, έχουν αναπτυχθεί οι κατάλληλες τεχνικές και τα εργαλεία ώστε η χρήσιμη πληροφορία που υπάρχει στα δημοφιλέστερα μέσα έκφρασης των καταναλωτών (διαδικτυακές πηγές και κοινωνικά μέσα) να μπορεί να αντληθεί και να υποστεί την κατάλληλη επεξεργασία και ανάλυση. Τα συμπεράσματα που προκύπτουν, οδηγούν στην καλύτερη κατανόηση των αναγκών/προσδοκιών των πελατών και τον εντοπισμό ή και την πρόβλεψη των τάσεων της αγοράς έτσι όπως τις διαμορφώνουν οι ίδιοι οι καταναλωτές.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, μελετάται η τεχνολογική πρόοδος στον χώρο της ανάλυσης της αγοράς και των τάσεων της και προτείνεται μια μεθοδολογία για την ανίχνευση των τάσεων της αγοράς για ένα συγκεκριμένο βιομηχανικό τομέα. Ο τομέας που επιλέχθηκε είναι αυτός της κατασκευής παιχνιδιών (toy manufacturing). Στόχος είναι μετά την πραγματοποίηση της έρευνας να επιλεχθεί ένας αποτελεσματικός συνδυασμός σύγχρονων τεχνικών για τον εντοπισμό των τάσεων της αγοράς στο συγκεκριμένο κλάδο μέσα από τα κοινωνικά δίκτυα και άλλες διαδικτυακές πηγές. Από το συνδυασμό των τεχνικών θα μπορεί να σχεδιαστεί ένα ολοκληρωμένο σύστημα ανίχνευσης τάσεων, το οποίο μπορεί να χρησιμεύσει σαν ένα εργαλείο επιχειρηματικής ευφυΐας στα χέρια μια εταιρίας αυτού του τομέα για την αναγνώριση πιθανών «κενών» ή ευκαιριών της εξεταζόμενης αγοράς, προς τα οποία μπορεί να προσανατολιστεί.

1.2. Δομή της διπλωματικής εργασίας

Στο παρόν κεφάλαιο (Κεφ.1, Εισαγωγή) αποτυπώνεται ο σκοπός της διπλωματικής εργασίας και η δομή της. Στη συνέχεια (Κεφ.2, Έρευνα Αγοράς και προτιμήσεων καταναλωτή), εξετάζεται σε θεωρητικό επίπεδο η Έρευνα Αγοράς, κυρίως στο ποσοστό που εμπεριέχει την έρευνα σχετικά με τις επιθυμίες και τις προτιμήσεις των καταναλωτών – με άλλα λόγια η ανίχνευση των τάσεων της αγοράς σαν απαραίτητη γνώση για την επιχείρηση από την σκοπιά του κλάδου του μάρκετινγκ καθώς και πώς αυτή πραγματοποιείται. Επίσης, δίνεται μια γενική εικόνα για τις σύγχρονες κατευθύνσεις στον τομέα της έρευνας αγοράς, ως μικρή εισαγωγή του τρίτου μέρους.

Στο τρίτο μέρος της εργασίας (Κεφ.3, Ανίχνευση των τάσεων της αγοράς από διαδικτυακές πηγές) γίνεται μια εισαγωγή στις πλέον σύγχρονες μεθόδους με τις οποίες μπορεί μια επιχείρηση να αντλήσει χρήσιμη πληροφορία από την ανεξάντλητη πηγή του διαδικτύου και των κοινωνικών μέσων για να βελτιστοποιήσει τις λειτουργίες της. Η διαδικασία αυτή είναι γνωστή σαν Επιχειρηματική Ευφυΐα (Business Intelligence), και οι μέθοδοι οι οποίες έχουν αναπτυχθεί τα τελευταία χρόνια για να μπορεί να αντληθεί αποτελεσματικά η πληροφορία από τον τεράστιο και συνεχώς ανανεούμενο όγκο δεδομένων του διαδικτύου (Big Data), ανήκουν στο χώρο της Εξόρυξης Δεδομένων (Data Mining). Στο κυρίως μέρος του κεφαλαίου, εξειδικεύουμε σε μια βιβλιογραφική επισκόπηση των μεθόδων που αφορούν το θέμα μας, δηλαδή την κατασκευή συστημάτων Επιχειρηματικής Ευφυΐας

βασισμένων σε μεθόδους Εξόρυξης Δεδομένων, ειδικά όμως για την ανίχνευση των τάσεων της αγοράς.

Στο τελευταίο μέρος της εργασίας (Κεφ.4, Ανίχνευση τάσεων της αγοράς στη βιομηχανία κατασκευής παιχνιδιών), γίνεται μια σύντομη παρουσίαση του κλάδου της κατασκευής παιχνιδιών και αναφορά στα ιδιαίτερα χαρακτηριστικά του σε θέματα μάρκετινγκ και έρευνας αγοράς. Εξετάζεται κατά πόσο είναι χρήσιμη η εξόρυξη δεδομένων από τα κοινωνικά δίκτυα και το διαδίκτυο για τον συγκεκριμένο κλάδο, και τέλος προτείνεται ένα ολοκληρωμένο σύστημα ανίχνευσης των τάσεων της αγοράς από τις πηγές αυτές για το συγκεκριμένο κλάδο.

2. Έρευνα Αγοράς και προτιμήσεων καταναλωτή

2.1. Έρευνα Αγοράς

Η Έρευνα Αγοράς αποτελεί τη συστηματική συλλογή και αξιολόγηση πληροφοριών από τις κατάλληλες ομάδες πληθυσμού με στόχο να βοηθήσει την επιχείρηση στη λήψη αποφάσεων με το μικρότερο δυνατό ρίσκο για υπάρχοντα και νέα προϊόντα και να ανακαλύψει νέες ευκαιρίες για τις δυνατότητες της επιχείρησης. Η αγορά είναι βασικός οικονομικός θεσμός, έχει να κάνει με την οικονομική συμπεριφορά των ανθρώπων, δηλαδή την διαδικασία παραγωγής και κατανάλωσης αγαθών [1]. Η έρευνα αγοράς είναι αναπόσπαστο κομμάτι της επιχειρηματικής στρατηγικής, καθώς με τη βοήθεια της έρευνας τα στελέχη μιας επιχείρησης μπορούν να βασιστούν σε αυτήν και να επιλύσουν σημαντικά στρατηγικά προβλήματα [2].

Με τον όρο «έρευνα αγοράς», εννοούμε «τη μελέτη σχετικά με το μέγεθος, τη σύνθεση και τα άλλα χαρακτηριστικά μιας συγκεκριμένης αγοράς» [3].

Μια ολοκληρωμένη έρευνα αγοράς, έχει σαν αντικειμενικό στόχο την παροχή πληροφοριών στη διοίκηση της επιχείρησης, αναφορικά με τον ακριβή προσδιορισμό του μεγέθους της αγοράς, την πρόβλεψη της ζήτησης για ολόκληρη την αγορά, την εκτίμηση της ζήτησης νέων αγαθών, την ανταγωνιστική θέση των προϊόντων της επιχείρησης, την ανάλυση των τοπικών δυνατοτήτων και χαρακτηριστικών [3].

Ο κύριος σκοπός της έρευνας είναι ο προσδιορισμός των προθέσεων των καταναλωτών για μια Επιχείρηση ή κάποιο προϊόν. Πέραν, αυτού όμως, η εν λόγω έρευνα παρουσιάζει αρκετά χρήσιμα στοιχεία που δείχνουν τις τάσεις της Αγοράς τόσο σε τοπικό επίπεδο όσο και σε επίπεδο χώρας, ή ακόμη και σε παγκόσμιο επίπεδο.

Η επιλογή της περιοχής βασίζεται σε διάφορες παραμέτρους που αφορούν την Επιχείρηση - Προϊον, όπως για παράδειγμα:

1. το target group που απευθύνεται η Επιχείρηση ή το προϊόν,
2. η κουλτούρα της Επιχείρησης,
3. τα εισοδήματα των καταναλωτών της περιοχής,
4. το ηλικιακό μίγμα του πληθυσμού της περιοχής, και
5. ο ανταγωνισμός που υφίσταται στη περιοχή.

2.2. Τα είδη της Έρευνας Αγοράς

Οι έρευνες αγοράς, που διεξάγονται στο φυσικό χώρο της αγοράς (field research), διακρίνονται σε κατηγορίες ανάλογα με:

- το σκοπό που επιδιώκουν,
- τη μεθοδολογία διεξαγωγής τους,
- το αντικείμενο της έρευνας,

- τον αριθμό των συμμετεχόντων στην έρευνα
- άλλα περισσότερο εξειδικευμένα κριτήρια που χρησιμοποιούν οι ερευνητές και αφορούν κυρίως στη μεθοδολογία που εφαρμόζουν για τη διεξαγωγή μιας έρευνας.

Οι χρήστες και οι ερευνητές στη μεταξύ τους επικοινωνία διακρίνουν την έρευνες αγοράς στις παρακάτω κατηγορίες.

1. Ποσοτικές και ποιοτικές έρευνες

Οι έρευνες αγοράς, ανάλογα με το μέγεθος του δείγματος και τη μεθοδολογία που διεξάγονται, διακρίνονται σε ποσοτικές και ποιοτικές. Οι ποσοτικές έρευνες διενεργούνται σε σχετικά μεγάλο και αντιπροσωπευτικό δείγμα του πληθυσμού και στοχεύουν στον ποσοτικό προσδιορισμό των σχέσεων των μεταβλητών του υπό διερεύνηση προβλήματος. Πρόκειται, δηλαδή, για περιγραφικές έρευνες που έχουν ως σκοπό να προσδιορίσουν ποιοι είναι αυτοί που αγοράζουν ένα συγκεκριμένο προϊόν ή κατηγορία προϊόντων, τι ποσότητες αγοράζουν, πόσο συχνά αγοράζουν. Οι ποιοτικές έρευνες είναι αυτές που διενεργούνται σε πολύ μικρές αντιπροσωπευτικές ομάδες του πληθυσμού και αποσκοπούν στον προσδιορισμό των ποιοτικών σχέσεων μεταξύ των μεταβλητών του προβλήματος. Πρόκειται δηλαδή για έρευνες που αποσκοπούν κυρίως στην αποκάλυψη των βαθύτερων αιτιών που καθορίζουν τη συμπεριφορά των ατόμων.

2. Έρευνες καταναλωτικές και βιομηχανίας ή εμπορίου

Οι έρευνες αγοράς, ανάλογα με το πεδίο που διενεργούνται, διακρίνονται σε καταναλωτικές έρευνες, βιομηχανίας και εμπορίου. Οι καταναλωτικές έρευνες διεξάγονται σε αντιπροσωπευτικά δείγματα του συνολικού πληθυσμού των καταναλωτών και στοχεύουν στη διερεύνηση των αγοραστικών συνηθειών τους, των στάσεων, των αντιλήψεων και απόψεών τους, σχετικά με την αγορά και κατανάλωση διαφόρων προϊόντων και υπηρεσιών, καθώς και τις συνήθειές τους σε ό,τι αφορά τη χρήση των Μ.Μ.Ε.

Οι έρευνες βιομηχανίας διεξάγονται σε αντιπροσωπευτικά δείγματα επιχειρήσεων ή οργανισμών και στοχεύουν στη συλλογή στοιχείων και πληροφοριών, σχετικά με τη χρήση ή τις προμήθειες διαφόρων προϊόντων ή υπηρεσιών, τη διερεύνηση των τρόπων συνεργασίας μεταξύ των διαφόρων επιχειρήσεων κ.ά. Οι έρευνες εμπορίου πραγματοποιούνται συνήθως σε σταθερό δείγμα καταστημάτων και στοχεύουν στη συλλογή στοιχείων και πληροφοριών, σχετικά με τη διακίνηση των διαφόρων προϊόντων στα σημεία πώλησης.

3. Έρευνες ad-hoc και έρευνες κοινής συμμετοχής

Οι έρευνες αγοράς, ανάλογα με το πλήθος των φορέων, για λογαριασμό των οποίων διεξάγεται η έρευνα διακρίνονται σε έρευνες αποκλειστικής συμμετοχής ή ad-hoc, όπως αποκαλούνται και σε έρευνες κοινής συμμετοχής. Οι έρευνες ad-hoc πραγματοποιούνται από τις εταιρίες ερευνών αποκλειστικά και μόνο για λογαριασμό ενός συγκεκριμένου φορέα, π.χ. μιας επιχείρησης. Τα αποτελέσματα αυτών των ερευνών δεν είναι ανακοινώσιμα σε τρίτους, εκτός αν το επιτρέψει αυτός που έδωσε εντολή για την διενέργεια της έρευνας. Οι έρευνες κοινής συμμετοχής διενεργούνται από εταιρίες ερευνών ή άλλους φορείς για λογαριασμό πολλών χρηστών που έχουν κοινά ερευνητικά ενδιαφέροντα. Ένα από τα κύρια πλεονεκτήματα, που παρουσιάζουν οι έρευνες αυτές, είναι η σχετικά χαμηλή τιμή τους, επειδή το συνολικό κόστος επιμερίζεται μεταξύ όλων των ενδιαφερομένων, γι' αυτό είναι και ιδιαίτερα ελκυστικές για τις μικρές και μεσαίες επιχειρήσεις που έχουν περιορισμένους πόρους. Χαρακτηριστικό παράδειγμα αποτελούν οι έρευνες Omnibus [4].

2.3. Παραδοσιακές μέθοδοι συλλογής στοιχείων

Οι ερευνητές και τα στελέχη του μάρκετινγκ τα οποία ασχολούνται με την Έρευνα Αγοράς, χρησιμοποιούν διάφορες μεθόδους προκειμένου να εντοπίσουν τις προτιμήσεις των καταναλωτών για κάποια χρονική περίοδο. Το ποια μέθοδος θα χρησιμοποιηθεί είναι προσαρμοσμένο στις ανάγκες της έρευνας, το έιδος της επιχείρησης, το είδος των προϊόντων που αφορά, ή ακόμα και τους διαθέσιμους πόρους για την διεξαγωγή της έρευνας.

Ο πιο άμεσος και αυτονόητος τρόπος για να συλλέξει το ενδιαφερόμενο μέρος δεδομένα από την σκοπιά των καταναλωτών είναι να τους ρωτήσει. Έτσι διακρίνονται οι παρακάτω άμεσες μέθοδοι έρευνας αγοράς οι οποίες αναλύονται συνοπτικά στη συνέχεια:

1. Προσωπικές συνεντεύξεις
2. Τηλεφωνική έρευνα
3. Ταχυδρομική έρευνα
4. Έρευνα μέσω Internet

1. Προσωπικές συνεντεύξεις

Η συλλογή των στοιχείων πραγματοποιείται μέσω προσωπικής συνέντευξης σε ένα ή περισσότερα δείγματα καταναλωτών, είτε με τη βοήθεια δομημένων ερωτηματολογίων, είτε με ελεύθερη συζήτηση. Στις ποσοτικές έρευνες η συλλογή των στοιχείων γίνεται πάντα με τη βοήθεια δομημένων ερωτηματολογίων. Στη διάρκεια αυτών των συνεντεύξεων ο ερευνητής υποβάλλει προς τον ερωτώμενο μια σειρά ανοικτών ερωτήσεων και καταγράφει τις απαντήσεις του. Στις ποιοτικές έρευνες οι συνεντεύξεις έχουν τη μορφή των προσωπικών συνεντεύξεων ή των ομαδικών συζητήσεων και γίνονται χωρίς τη χρήση ερωτηματολογίου. Συνήθως, την Έρευνα Αγοράς με συνεντεύξεις αναλαμβάνουν εξειδικευμένες εταιρείες για λογαριασμό κάποιας επιχείρησης, και όσοι συμμετέχουν σε αυτές πληρώνονται. Έτσι πρόκειται για διαδικασία που απαιτεί αρκετούς πόρους σε χρόνο και χρήμα για την επιχείρηση, μπορεί όμως αν πραγματοποιηθεί και αναλυθεί σωστά να οδηγήσει σε πολύ αξιόπιστα αποτελέσματα.

2. Τηλεφωνική έρευνα

Παρόμοια με την προσωπική συνέντευξη, στην τηλεφωνική έρευνα ο ερευνητής συλλέγει στοιχεία κάνοντας στοχευμένες ερωτήσεις σε ένα δείγμα της ομάδας στόχου του (target group) μέσω τηλεφώνου. Λόγω του απρόσωπου χαρακτήρα της τηλεφωνικής επικοινωνίας, χρειάζεται μια παραπάνω προσπάθεια ώστε το ενδιαφέρον του ερωτώμενου να κρατηθεί σταθερό και να απαντήσει σε όλες τις ερωτήσεις [5]. Για το λόγο αυτό οι ερευνητές πρέπει να είναι εξειδικευμένοι και κατάλληλα προετοιμασμένοι, να διαθέτουν πειθώ, ευγένεια και καλή άρθρωση. Η τηλεφωνική έρευνα μπορεί να χρησιμοποιηθεί για τη μέτρηση του βαθμού διείσδυσης ή της χρήσης διάφορων προϊόντων στα νοικοκυριά, για τη μέτρηση του βαθμού γνώσης τους (awareness), για έρευνες media (αναγνωσιμότητας, ακροαματικότητας κ.λπ.), κ.λπ. Μπορεί, επίσης, να χρησιμοποιηθεί για έρευνα σε ξένες αγορές, γιατί η δυνατότητα τηλεφωνικής επικοινωνίας μπορεί να εξασφαλίσει φτηνή και σύντομη συλλογή πληροφοριών [6].

3. Έρευνα μέσω ταχυδρομείου

Ως έρευνα μέσω ταχυδρομείου εννοούμε την ταχυδρομική αποστολή ερωτηματολογίων στους τελικούς καταναλωτές όπου θέλει να απευθυνθεί ο ερευνητής, ενώ έχει προβλεφθεί τρόπος για την επιστροφή του ερωτηματολογίου σε αυτόν. Πρόκειται για έναν οικονομικό και εύκολο τρόπο έρευνας, που όμως είναι ακόμα πιο επισφαλής από την τηλεφωνική επικοινωνία καθώς τίθεται εύλογα υπό αμφισβήτηση η αξιοπιστία των αποτελεσμάτων. Η χρήση της ταχυδρομικής έρευνας από άτομα που δεν διαθέτουν την κατάλληλη πείρα σ' αυτόν τον τομέα, μπορεί να δώσει αποτελέσματα άσχετα με το πραγματικό πρόβλημα, του οποίου επιδιώκεται η διερεύνηση, και κατά συνέπεια να οδηγήσει σε λανθασμένα συμπεράσματα και επικίνδυνες επιχειρηματικές αποφάσεις. Η επιτυχία μιας ταχυδρομικής έρευνας εξαρτάται κατά κύριο λόγο από βαθμό ανταπόκρισης των ερωτωμένων, επομένως απαιτείται ειδικός σχεδιασμός και επιλογή του δείγματος, ώστε να εξασφαλίζεται υψηλός βαθμός ανταπόκρισης. Επίσης, απαιτείται ειδικός σχεδιασμός του ερωτηματολογίου, στο οποίο, εκτός των ερωτήσεων και των προκωδικοποιημένων απαντήσεων, πρέπει να υπάρχουν και σαφείς οδηγίες συμπλήρωσης του ερωτηματολογίου προς τον ερωτώμενο. Αυτός ο τρόπος έρευνας συνήθως χρησιμοποιείται όταν: 1. υπάρχουν περιορισμένα κονδύλια για έρευνα που δεν επιτρέπουν τη διεξαγωγή ποσοτικών ερευνών με άλλο τρόπο 2. το «δείγμα» δεν είναι εύκολο να το προσεγγίσει ο ερευνητής, π.χ., στελέχη επιχειρήσεων που δύσκολα βρίσκουν χρόνο για προσωπική συνέντευξη ή 3. οι απαντήσεις στα ερωτήματα της έρευνας απαιτούν περισσότερη σκέψη και λιγότερο αυθορμητισμό [6].

4. Έρευνα μέσω Internet

Η τεράστια ανάπτυξη του διαδικτύου (Internet), όπως ήταν φυσικό, έχει επηρεάσει και την Έρευνα Αγοράς. Τα τελευταία χρόνια το Διαδίκτυο έχει αρχίσει να χρησιμοποιείται και για τη διεξαγωγή ερευνών αγοράς και προτιμήσεων καταναλωτή. Η Έρευνα Αγοράς μέσω Internet γίνεται με δυο τρόπους [7]:

1. Με προβολή ερωτηματολογίων σε ειδικές ιστοσελίδες, στα οποία καλούνται να συμπληρώσουν on line οι επισκέπτες των συγκεκριμένων sites
2. Με την αποστολή ερωτηματολογίων μέσω ηλεκτρονικού ταχυδρομείου (e-mail)

Ωστόσο, παρά την ευκολία που υπάρχει στη συλλογή των στοιχείων μέσω του Internet, η χρήση του διαδικτύου για τη διεξαγωγή ερευνών είναι ακόμα εξαιρετικά περιορισμένη, λόγω ειδικών προβλημάτων που υπάρχουν και περιορίζεται κυρίως σε ειδικές εφαρμογές, όπως την έρευνα μεταξύ χρηστών προγραμμάτων λογισμικού [8].

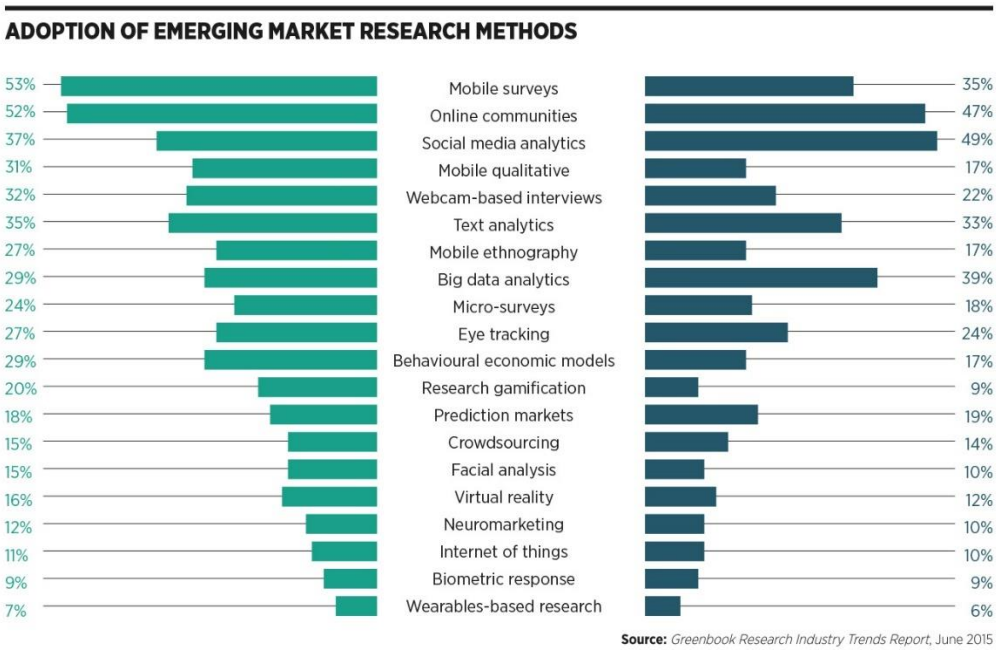
2.4. Σύγχρονες κατευθύνσεις

Ενώ οι παραδοσιακές μέθοδοι συλλογής στοιχείων προσαρμοσμένων για σκοπούς έρευνας αγοράς συνεχίζουν να υπάρχουν και να αποτελούν αντικείμενο εξειδικευμένων εταιρειών, από τη δεκαετία του 2000 άρχισε να αναπτύσσεται περισσότερο η διαδικτυακή (online) πλευρά της συλλογής δεδομένων. Η ανάπτυξη της διαδικτυακής έρευνας οδήγησε σε βελτιωμένη ταχύτητα και μειωμένο κόστος για τη συλλογή των δεδομένων.

Ενδεικτικά αναφέρουμε ότι το 2005, η Αυστραλία γίνεται η πρώτη χώρα όπου η συλλογή δεδομένων από το διαδίκτυο είναι ο δημοφιλέστερος τρόπος συλλογής πληροφοριών, ενώ το

λανσάρισμα του iphone το 2007 από την Apple οδήγησε στις πρώτες εφαρμογές συλλογής δεδομένων βασισμένες στη λειτουργία του smartphone. Σταδιακά, η έρευνα μέσω κινητού (mobile research) έγινε ιδιαίτερα δημοφιλής, όπως και η έρευνα μέσω των κοινωνικών μέσων (social media research), ενώ ο όρος Μεγάλα Δεδομένα (Big Data) άρχισε να αναφέρεται με όλο και μεγαλύτερη συχνότητα σε κάθε κλάδο σχετιζόμενο με τη συλλογή πληροφορίας προερχόμενης από το ευρύ κοινό.

Η ολοένα και αυξανόμενη χρήση του διαδικτύου και των εφαρμογών για κινητά, και ιδιαίτερα των κοινωνικών μέσων, έχει δώσει στους καταναλωτές ένα μέσο αυθόρμητης έκφρασης προτιμήσεων και απόψεων, και στις επιχειρήσεις μια ανεξάντλητη πηγή πληροφορίας σχετικής με τις τρέχουσες τάσεις της αγοράς σε πραγματικό χρόνο. Σε συνδυασμό με την τεχνολογική πρόοδο στους τομείς της Επιχειρηματικής Ευφυΐας, της Εξόρυξης Δεδομένων και των Μεγάλων δεδομένων, οι επιχειρήσεις πλέον έχουν τις πηγές αλλά και τα μέσα για να έχουν μια ολοκληρωμένη εικόνα της αγοράς και των τάσεών της μέσα από τα μάτια των καταναλωτών. Οι τομείς αυτοί αποτελούν τις σύγχρονες κατευθύνσεις για την έρευνα αγοράς, και αναλύονται εκτενώς στο επόμενο κεφάλαιο. Στην Εικόνα 1, παρουσιάζεται το ποσοστό υιοθέτησης καινοτόμων μεθόδων όπως οι έρευνες εφαρμογών σε κινητά, η διερεύνηση διαδικτυακών κοινοτήτων και η εξόρυξη δεδομένων από μέσα κοινωνικής δικτύωσης για την Έρευνα Αγοράς.



Εικόνα 1: Ποσοστό υιοθέτησης ανερχόμενων μεθόδων Έρευνας Αγοράς

3. Ανίχνευση Τάσεων Αγοράς από διαδικτυακές πηγές

3.1. Επιχειρηματική Ευφυΐα

3.1.1. Εισαγωγή

Προκειμένου μια επιχείρηση να διασφαλίσει τη βιωσιμότητα και την ανταγωνιστικότητα στον κλάδο δραστηριοποίησής της, είναι απαραίτητο να διαμορφώνει τη στρατηγική και το σχεδιασμό της ανάλογα με τα τρέχοντα δεδομένα της αγοράς. Για το σκοπό αυτό οφείλει να μπορεί να χρησιμοποιεί τις κατάλληλες τεχνικές και μεθοδολογίες προκειμένου να συγκεντρώνει και να επεξεργάζεται τα δεδομένα που την αφορούν, αλλά και να τα μετατρέπει σε χρήσιμες για τους σκοπούς της πληροφορίες. Η διαδικασία αυτή αναφέρεται στη βιβλιογραφία ως Επιχειρηματική Ευφυΐα (business intelligence), και αποτελεί ένα υποστηρικτικό εργαλείο για τη λήψη αποφάσεων μέσα στην επιχείρηση (decision making). Η ανταγωνιστική ευφυΐα (competitive intelligence) όπως και η ευφυΐα της αγοράς (marketing intelligence) αποτελούν υποκατηγορίες της Επιχειρηματικής Ευφυΐας, προσανατολισμένες στα δεδομένα σε σχέση με τις δράσεις των ανταγωνιστών και τις προτιμήσεις των καταναλωτών αντίστοιχα.

Ένα ολοκληρωμένο **σύστημα Επιχειρηματικής Ευφυΐας** είναι σχεδιασμένο για να καλύπτει τέσσερις ανάγκες της επιχείρησης:

1. Να αναγνωρίζει τις ευκαιρίες και τις απειλές που ενυπάρχουν στο περιβάλλον της αγοράς
2. Να παρέχει πληροφορίες σχετικά με τον ανταγωνισμό
3. Να βοηθά στην πρόβλεψη των μελλοντικών δράσεων των ανταγωνιστών
4. Να βοηθά στην αποτελεσματική λήψη αποφάσεων στον τομέα του marketing. [9]

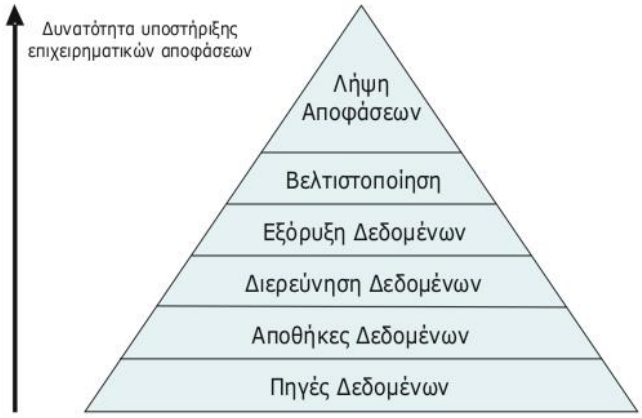
Πιθανά οφέλη από την ενσωμάτωση εργαλείων Επιχειρηματικής Ευφυΐας στην επιχειρησιακή λειτουργία είναι η επιτάχυνση και η βελτίωση της λήψης αποφάσεων, η βελτιστοποίηση των εσωτερικών διαδικασιών, η αύξηση της λειτουργικής αποτελεσματικότητας, η δημιουργία μεγαλύτερων εσόδων αλλά και ανταγωνιστικού πλεονεκτήματος. Τα συστήματα Επιχειρηματικής Ευφυΐας (BI) μπορούν επίσης να βοηθήσουν τις εταιρείες να αναγνωρίσουν τις τάσεις της αγοράς και να εντοπίσουν προβλήματα που χρειάζονται διευθέτηση.

Αρχικά, τα εν λόγω συστήματα χρησιμοποιούνταν από αναλυτές δεδομένων (data analysts) και άλλους επαγγελματίες της πληροφορικής (IT professionals) οι οποίοι «έτρεχαν» τα κατάλληλα προγράμματα και παρήγαγαν αναφορές με τα αποτελέσματα προς τους μη-εξειδικευμένους χρήστες (business users). Σταδιακά τα στελέχη και οι υπάλληλοι των επιχειρήσεων μπόρεσαν να χρησιμοποιήσουν το λογισμικό της επιχειρηματικής ανάλυσης χωρίς τη διαμεσολάβηση ειδικού, χάρη στην ανάπτυξη του τομέα και τη δημιουργία εργαλείων υψηλού επιπέδου φιλικών προς το μέσο χρήστη.

Η Επιχειρηματική Ευφυΐα σαν όρος χρησιμοποιείται συνήθως με τον ίδιο τρόπο όπως και η επιχειρηματική ανάλυση. Σε κάποιες περιπτώσεις, η επιχειρηματική ανάλυση αναφέρεται υπονοώντας το πιο στενό πλαίσιο των προηγμένων μεθόδων για ανάλυση δεδομένων (advanced data analytics) η σαν γενικότερη έννοια που περιλαμβάνει τόσο το BI όσο και την ανάλυση δεδομένων.

Σε αυτό το σημείο μπορεί να δοθεί μια πρώτη εικόνα της δομής των συστημάτων Επιχειρηματικής Ευφυΐας, καθώς παρά τη μεγάλη ποικιλία σε εργαλεία και μεθοδολογίες τα στάδια της διαδικασίας

είναι παραπλήσια: πρόκειται για τα βήματα που οδηγούν από την αδόμητη πληροφορία σε πραγματική, επιχειρηματικά ή επιχειρησιακά χρήσιμη γνώση. Στη βάση της δομής βρίσκονται τα αρχικά ακατέργαστα δεδομένα, ενώ στην κορυφή της βρίσκεται η λήψη των τελικών αποφάσεων. Κάθε μετάβαση από ένα επίπεδο σε κάποιο ανώτερο, αυξάνει τη δυνατότητα υποστήριξης επιχειρηματικών αποφάσεων.



Εικόνα 2: Η πυραμίδα συστημάτων Επιχειρηματικής Ευφυΐας [10]

3.1.2. Ιστορία

Σποραδική χρήση του όρου «Επιχειρηματική Ευφυΐα» γινόταν τουλάχιστον από τη δεκαετία του 1860, ενώ επίσημη χρήση του καθιερώθηκε από το 1989, σαν μια κατηγορία-ομπρέλα για τις εφαρμογές και τις τεχνικές ανάλυσης δεδομένων που υποστηρίζουν τις διαδικασίες λήψης αποφάσεων. Οι τεχνολογίες που αναπτύχθηκαν ξεκίνησαν από την εποχή των πρώτων υπολογιστικών συστημάτων (mainframe/legacy systems), πάνω στα οποία άρχισε να γίνεται η εξόρυξη και ανάλυση της πληροφορίας.

Τα πρώτα συστήματα Επιχειρηματικής Ευφυΐας αναπτύχθηκαν από την IBM και τη Siebel (η οποία απορροφήθηκε από την Oracle) την περίοδο μεταξύ 1970 και 1990. Το 1988, ειδικοί της τεχνολογίας λογισμικού και επενδυτές διοργάνωσαν το πρώτο συνέδριο με θέμα την ανάλυση δεδομένων (Multiway Data Analysis Consortium) στη Ρώμη, όπου συζητήθηκε πώς μπορεί να γίνει η διαχείριση και ανάλυση των δεδομένων πιο αποδοτική, και προσιτή σε μικρότερες και οικονομικά πιο αδύναμες επιχειρήσεις. Μέχρι το 2000 είχαν ήδη αναπτυχθεί πολυάριθμα συστήματα δημιουργίας αναφορών και λογισμικά ανάλυσης, κάποια εκ των οποίων από τους μεγαλύτερους παραγωγούς λογισμικού των ΗΠΑ.

Τη δεκαετία του 2000, το ενδιαφέρον φαίνεται να επικεντρώθηκε στη δημιουργία BI συστημάτων υψηλής συμβατότητας σε διαφορετικά περιβάλλοντα και με μικρό κόστος εγκατάστασης. Αυτές οι φιλοδοξίες σε συνδυασμό με την τάση για μεταφορά των δεδομένων και των προγραμμάτων σε cloud, ώθησαν στην ανάπτυξη ανεξάρτητων συστημάτων με απεριόριστη πρόσβαση στην πληροφορία. Τα θετικά αποτελέσματα της αποθήκευσης σε cloud φάνηκαν περισσότερο από το 2006 και μετά, καθώς άρχισε να εδραιώνεται η απομακρυσμένη πρόσβαση στην πληροφορία από οποιαδήποτε τοποθεσία και μέσα από διαφορετικές συσκευές και προγράμματα περιήγησης.

Η ανάλυση δεδομένων σε cloud που επέτρεψε επίσης την επεξεργασία μεγάλου όγκου πληροφορίας,

η πρόοδος στο υπολογιστικό υλικό σε θέματα μνήμης και ταχύτητας καθώς και η γενικότερη ανεμπόδιστη τεχνολογική εξέλιξη, συνέβαλλαν σημαντικά στο πεδίο της ανάπτυξης λογισμικού Επιχειρηματικής Ευφυΐας/ανάλυσης και την επόμενη δεκαετία. Στις αρχές του 2016 η αγορά των σχετικών προϊόντων έφτασε προσεγγιστικά τα 9 δις. δολλάρια, με τη λειτουργικότητά τους να ξεπερνά κατά πολύ την απλή ανάλυση των δεδομένων. Οι σημερινές εφαρμογές έχουν τη δυνατότητα να λύνουν προβλήματα marketing, να διεξάγουν λεπτομερείς διαγνώσεις σχετικά με τη λειτουργία της επιχείρησης και να λειτουργούν με προσαρμοστικότητα σε διαφορετικά περιβάλλοντα και επιχειρηματικά οικοσυστήματα.

3.1.3. Εργαλεία

Στην κατηγορία των BI συστημάτων εντάσσεται μια μεγάλη ποικιλία εργαλείων, εφαρμογών και μεθοδολογιών που επιτρέπει στις επιχειρήσεις να συλλέγουν δεδομένα από εσωτερικές και εξωτερικές πηγές, να τα προετοιμάζουν για ανάλυση, να διατυπώνουν και να τρέχουν ερωτήματα στις βάσεις των δεδομένων και να δημιουργούν αναφορές (reports), πίνακες και οπτικοποιημένα αποτελέσματα που διευκολύνουν την επιχειρηματική ανάλυση και λειτουργικότητα.

Ανάμεσα στο ευρύ σύνολο των καινοτόμων σύγχρονων εφαρμογών, συγκαταλέγονται εργαλεία ad-hoc ανάλυσης και αναζήτησης, εργαλεία δημιουργίας αναφορών, ανάλυσης σε πραγματικό χρόνο (online analytical processing – OLAP), εφαρμογές για κινητά (mobile BI), εργαλεία πραγματικού χρόνου (real-time BI), προγράμματα και λογισμικό σε cloud, εργαλεία συνδιαχείρισης (collaborative BI) κ.α. Επίσης στη γενικότερη κατηγορία εντάσσεται το λογισμικό για την οπτικοποίηση των δεδομένων (σχεδιασμός πινάκων και διαγραμμάτων) όπως επίσης και τα εργαλεία για τη δημιουργία περιβάλλοντος με σαφή λειτουργικότητα φιλικό προς τον τελικό χρήστη (dashboards, interfaces). Όσες εφαρμογές είναι απαραίτητες μπορούν να αγοραστούν είτε σαν ενοποιημένη πλατφόρμα από έναν προμηθευτή είτε από διαφορετικούς προμηθευτές-κατασκευαστές.

Τα συστήματα BI μπορούν επίσης να περιέχουν μορφές προχωρημένης ανάλυσης δεδομένων, όπως είναι η **Εξόρυξη Δεδομένων** (data mining), η ανάλυση με σκοπό την πρόβλεψη της μελλοντικής συμπεριφοράς των δεδομένων, η ανάλυση κειμένου και φυσικής γλώσσας, και η ανάλυση μεγάλου όγκου πληροφορίας (big data analytics). Σε πολλές περιπτώσεις αυτή η πιο εξειδικευμένη μορφή ανάλυσης γίνεται ακόμα από ομάδες ειδικών από διαφορετικούς κλάδους και δεν παρέχεται εύκολα από τα δεδομένα εξόδου ολοκληρωμένων εμπορικών εργαλείων λόγω αυξημένης πολυπλοκότητας.

3.2. Εξόρυξη πληροφορίας από δεδομένα

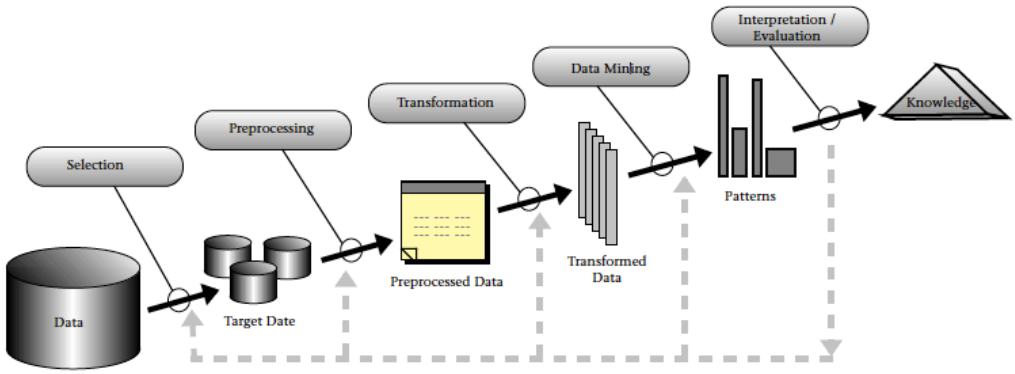
3.2.1. Γενικά

Όλες οι διαδικασίες, οι τεχνικές και τα εργαλεία με τα οποία γίνεται συγκέντρωση και ανάλυση της πληροφορίας από τα ενδιαφερόμενα μέρη, αποτελούν πεδίο έρευνας του κλάδου της εξόρυξης δεδομένων, ή εξόρυξης πληροφορίας από δεδομένα (data mining). Η εξόρυξη δεδομένων είναι ένας διεπιστημονικός κλάδος της επιστήμης των υπολογιστών. Αφορά την υπολογιστική διαδικασία του εντοπισμού μοτίβων σε ένα μεγάλο σύνολο δεδομένων, και εμπλέκει μεθόδους τεχνητής νοημοσύνης (artificial intelligence), μηχανικής μάθησης (machine learning), στατιστικής και βάσεων δεδομένων. Η πρόοδος που έχει γίνει στον κλάδο της εξόρυξης δεδομένων τα τελευταία χρόνια ήταν

απαραίτητη για τον εντοπισμό χρήσιμων πληροφοριών μέσα στον τεράστιο όγκο δεδομένων που προσφέρουν οι πλατφόρμες των μέσων κοινωνικής δικτύωσης. Συγκεκριμένα, επιτρέπει τον εντοπισμό τάσεων, μοτίβων και κανόνων που διέπουν την πληροφορία που βρίσκεται στο διαδίκτυο. Οι τεχνικές που χρησιμοποιούνται αφορούν την επεξεργασία των δεδομένων, την ανάλυσή τους και σε τελευταίο στάδιο την ερμηνεία τους. [11]

Ιστορικά, η διαδικασία της εύρεσης χρήσιμων μοτίβων στα δεδομένα έχει εκφραστεί με διάφορους τρόπους, όπως εξόρυξη δεδομένων, εξαγωγή γνώσης (knowledge extraction), ανακάλυψη γνώσης (information discovery), συγκομιδή γνώσης (information harvesting) και επεξεργασία μοτίβων δεδομένων (data pattern processing) [12]. Ο όρος εξόρυξη δεδομένων αρχικά χρησιμοποιήθηκε περισσότερο από στατιστικολόγους, αναλυτές δεδομένων και τις κοινότητες των συστημάτων διαχείρισης πληροφοριών (management information systems - MIS). Κέρδισε επίσης δημοτικότητα στο πεδίο των βάσεων δεδομένων. Αντίστοιχα, ο όρος εξαγωγή γνώσης χρησιμοποιήθηκε συχνότερα στα πεδία της τεχνητής νοημοσύνης και της μηχανικής μάθησης.

Στο ευρύτερο πλαίσιο της εξαγωγής γνώσης σε ένα σύστημα Επιχειρηματικής Ευφυΐας συγκεκριμένης λειτουργικότητας, η εξόρυξη δεδομένων μπορεί να θεωρηθεί ως τμήμα αυτής, μαζί με τη συγκέντρωση των δεδομένων, την προεπεξεργασία τους και την αξιολόγηση των συμπερασμάτων.



Εικόνα 3: Η Εξόρυξη Δεδομένων ως τμήμα της διαδικασίας Εξαγωγής Γνώσης [12]

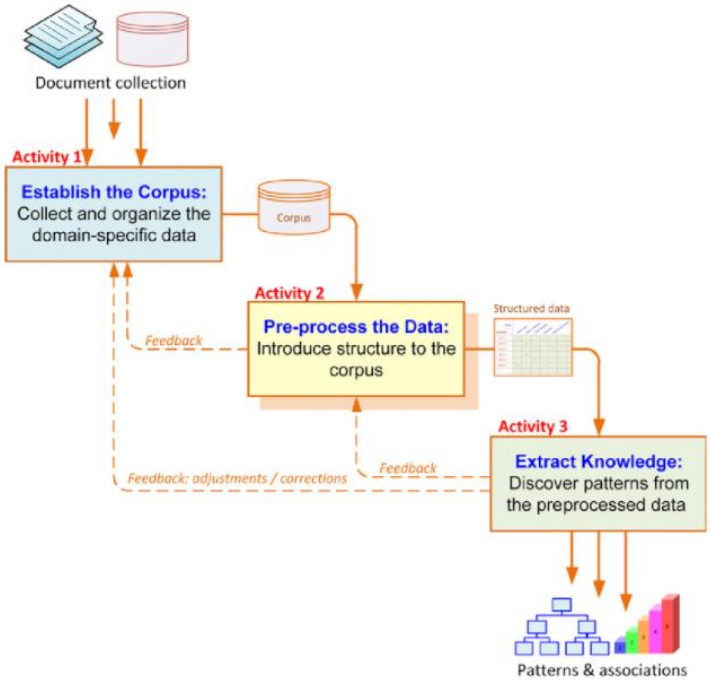
3.2.2. Τεχνικές Εξόρυξης Δεδομένων

Ανάμεσα στις πιο γνωστές τεχνικές εξόρυξης δεδομένων, με εφαρμογές για πολλούς διαφορετικούς σκοπούς, είναι:

1. Οι Στατιστικές μέθοδοι (π.χ. γραμμικές μέθοδοι πρόβλεψης, ιστογράμματα)
2. Οι Αλγόριθμοι εύρεσης πλησιέστερου γείτονα (NN, k-NN)
3. Η Κατηγοριοποίηση (classification), η οποία χρησιμοποιείται από μεθόδους που επιτρέπουν την πρόβλεψη της κατηγορίας στην οποία ανήκει ένα αντικείμενο με βάση τα χαρακτηριστικά του
4. Η Ομαδοποίηση/Ανάλυση Συστάδων (clustering), η οποία επιτρέπει τον εντοπισμό ομάδων ομοειδών αντικειμένων

5. Οι Μέθοδοι βασισμένες σε Δέντρα και Γράφους
6. Τα Νευρωνικά Δίκτυα
7. Η Εφαρμογή Κανόνων Συσχέτισης

Το στάδιο της εξόρυξης δεδομένων, όπως έχει αναφερθεί και νωρίτερα, αφορά στον εντοπισμό προτύπων μέσα σε ένα σύνολο δεδομένων που έχει ήδη συγκεντρωθεί και στις περισσότερες περιπτώσεις έχει υποστεί και μια προ-επεξεργασία. Ανάλογα με το είδος της ανάλυσης που διεξάγεται, τις απαιτήσεις και το γενικότερο πλαίσιο της μεθοδολογίας που χρησιμοποιείται, επιλέγεται και ο τρόπος με τον οποίο γίνεται η εξόρυξη των δεδομένων. Αυτό μπορεί να σημαίνει είτε κάποια από τις δημοφιλείς τεχνικές είτε συνδυασμό δύο ή περισσότερων εξ αυτών, είτε κάποια καινοτόμα πρόταση καθώς πρόκειται για ένα πεδίο συνεχόμενης επιστημονικής έρευνας με αυξημένο ενδιαφέρον.



Εικόνα 4: Η Εξόρυξη Δεδομένων από κείμενο σαν διαδικασία τριών σταδίων [13]

Ένας βασικός διαχωρισμός των μεθόδων εξόρυξης δεδομένων είναι σε μεθόδους επιβλεπόμενης μάθησης (supervised learning) και μεθόδους μη επιβλεπόμενης μάθησης (unsupervised learning) [10]. Η επιβλεπόμενη μάθηση έχει στόχο τη μοντελοποίηση των σχέσεων ανάμεσα σε ένα εξαρτημένο γνώρισμα – στόχο και σε άλλα ανεξάρτητα γνωρίσματα. Η ανάλυση συνίσταται στην τυποποίηση των σχέσεων ανάμεσα στην εξαρτημένη και στις ανεξάρτητες μεταβλητές, συνήθως με τη δημιουργία ενός μοντέλου, που επιτρέπει τον υπολογισμό της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Το μοντέλο μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Το όνομα «επιβλεπόμενη μάθηση» σημαίνει ότι το γνώρισμα στόχος και οι τιμές του καθοδηγούν τη διαδικασία μάθησης. Στη μη επιβλεπόμενη μάθηση δεν υπάρχει κάποια στήλη στόχος και οι αλγόριθμοι προσπαθούν να ομαδοποιήσουν τα δεδομένα σε ομάδες που δεν είναι γνωστές εκ των προτέρων.

3.2.3. Δεδομένα μεγάλου όγκου

Η έννοια των δεδομένων μεγάλου όγκου ή Μεγάλων Δεδομένων (Big Data) ξεκίνησε να χρησιμοποιείται στα πλαίσια της επιστήμης των υπολογιστών ήδη από το ξεκίνημά της. Στην αρχή αναφερόταν σε ένα σύνολο δεδομένων το οποίο είναι αρκετά μεγάλο ώστε να μη μπορεί να υποστεί αποδοτική επεξεργασία λόγω ανεπάρκειας των παραδοσιακών μεθόδων και εργαλείων. Η πρώτη εμφάνιση του όρου, φέρεται να έγινε το 1997 από τους επιστήμονες της NASA: ανέφεραν ότι αδυνατούσαν να αναπαραστήσουν γραφικά (visualization) τα σύνολα δεδομένων που κατείχαν (data sets), καθώς ήταν τόσο μεγάλα που ήταν ακατόρθωτο να τα αποθηκεύσουν στη κύρια μνήμη, στον τοπικό δίσκο και σε εξωτερικό σκληρό δίσκο. Έτσι δήλωσαν ότι αντιμετωπίζουν πρόβλημα Μεγάλων Δεδομένων.

Με την πάροδο του χρόνου και την πρόοδο του υπολογιστικού υλικού (hardware) αποθήκευσης, ο όρος αντιπροσώπευε δεδομένα ολοένα και μεγαλύτερα σε όγκο. Υπολογίζεται ότι σήμερα με το συγκεκριμένο όρο αναφερόμαστε συνήθως σε όγκους δεδομένων που κυμαίνονται από μερικά terabytes έως δεκάδες ή και εκατοντάδες petabytes (1,024 terabytes) ή exabytes (1,024 petabytes) ή zetabytes (1,024 exabytes).

Αρχικά ο όρος Μεγάλα Δεδομένα χρησιμοποιούνταν κυρίως για δομημένα δεδομένα (structured data), τα οποία είχε στη διάθεσή του το ενδιαφερόμενο μέρος από μια συγκεκριμένη πηγή και για συγκεκριμένο σκοπό και απλά αδυνατούσε να τα επεξεργαστεί ή/και να τα αποθηκεύσει. Αυτό είναι κάτι που άλλαξε ριζικά στην πορεία, αφού η δυναμική της χρησιμοποίησης των νέων τεχνολογιών έδειξε πως η περισσότερη πληροφορία σε παγκόσμιο επίπεδο δημιουργείται και διαδίδεται μαζικά, με αδόμητο τρόπο και με ποικιλία στις μορφές της (κείμενο, πολυμέσα).

Τα μέσα κοινωνικής δικτύωσης αποτελούν ένα από τα πιο σύγχρονα παραδείγματα πηγών δεδομένων μεγάλου όγκου, όπως και οι υπόλοιπες διαδικτυακές πηγές που ανανεώνουν με μεγάλη συχνότητα το περιεχόμενό τους. Ενδεικτικά, κάθε ημέρα κατά μέσο όρο:

1. Στέλνονται 294 δις. μηνύματα ηλεκτρονικού ταχυδρομείου
2. Πραγματοποιούνται πάνω από 1 δις. αναζητήσεις στο Google
3. Δημοσιεύονται πάνω από 230 εκ. αναρτήσεις στο Twitter,

ενώ στο Facebook υπάρχουν αποθηκευμένα, προσβάσιμα και έτοιμα για ανάλυση πάνω από 30 Petabytes δεδομένα δημιουργημένα από χρήστες (*Πηγή: IBM Big Data & Analytics Hub).*

Πέρα από τον κόσμο της πληροφορίας του διαδικτύου, τα Μεγάλα Δεδομένα αποτελούν πρόκληση για πολλές βιομηχανίες και τομείς, όπως η επιστημονική έρευνα, οι επικοινωνίες, η υγεία, η εκπαίδευση, η οικονομία, κ.α.

Σε σχέση με τα χαρακτηριστικά τους, τα δεδομένα μεγάλου όγκου έχουν εκφραστεί μέσα από το μοντέλο των 3V, το οποίο ορίστηκε το 2001 από τον Douglas Laney [14] ως *«Μεγάλος όγκος, μεγάλη ταχύτητα και μεγάλη ποικιλία στην πληροφορία, η οποία απαιτεί καινοτόμες, οικονομικές σε πόρους μεθόδους για την επεξεργασία πληροφορίας, ώστε να ενεργοποιηθούν όλες οι δυνατότητες για λήψη αποφάσεων και αυξημένη διορατικότητα»*. Το 2012 ο αρχικός ορισμός επαναδιατυπώθηκε ως εξής: *«Τα Μεγάλα Δεδομένα είναι στοιχεία πληροφορίας με μεγάλο όγκο, μεγάλη ταχύτητα και/ή μεγάλη ποικιλία τα οποία απαιτούν για την επεξεργασία τους νέες μεθόδους ώστε να ενεργοποιηθούν όλες οι δυνατότητες για λήψη αποφάσεων, αυξημένη διορατικότητα και βελτιστοποίηση των*

διαδικασιών». Σύμφωνα με τον ορισμό, τα τρία βασικά χαρακτηριστικά των δεδομένων μεγάλου όγκου σύμφωνα με το μοντέλο 3V είναι Όγκος, Ταχύτητα και Ποικιλία (**V**olume, **V**elocity, **V**ariety):

1. Ο όρος **Όγκος (Volume)** αναφέρεται στην ποσότητα των δεδομένων του εκάστοτε συνόλου προς ανάλυση, ο οποίος μπορεί να καλύπτει μέχρι και Terabytes ή ακόμα και Petabytes αποθηκευτικού χώρου.

2. Ο όρος **Ποικιλία (Variety)** αναφέρεται στο μεγάλο εύρος πιθανών διαφορετικών τύπων δεδομένων προς ανάλυση. Ο όρος ανταποκρίνεται στην ανάγκη να καταχωρούνται, να επεξεργάζονται και να συνδυάζονται δεδομένα διαφορετικών πηγών, κάτι που μπορεί να σημαίνει όχι μόνο διαφορετικούς τύπους δεδομένων, αλλά και διαφορετική δομή μεταξύ των ιδίων τύπων. Σε πρώτη φάση, δημιουργείται έτσι η ανάγκη να ενσωματωθούν δεδομένα αυστηρώς δομημένα (structured), ημιδομημένα (semi-structured) και αδόμητα (unstructured).

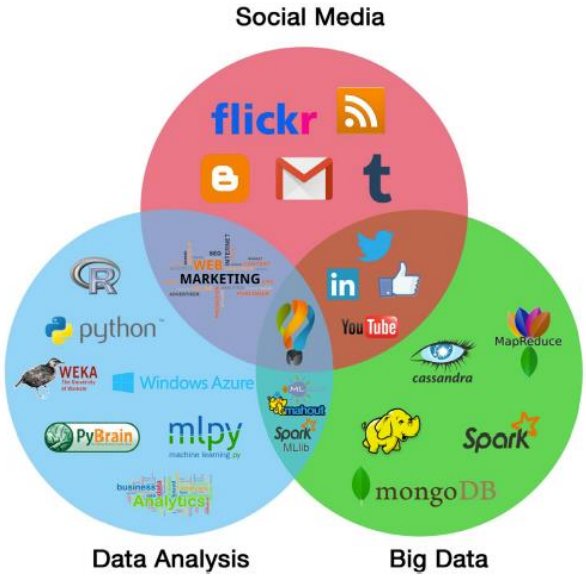
3. Ο όρος **Ταχύτητα (Velocity)** αναφέρεται στον ταχύτατο ρυθμό με τον οποίο εισέρχονται νέα δεδομένα στο εκάστοτε σύστημα αλλά και ανανεώνονται τα ήδη υπάρχοντα. Επιπλέον, έχει να κάνει με τον χρόνο που απαιτείται για την επεξεργασία και ανάλυση τους κατά την είσοδό τους στο σύστημα.

Η σημασία των Μεγάλων Δεδομένων για τον επιχειρηματικό κόσμο είναι τεράστια, καθώς τα στελέχη των επιχειρήσεων μπορούν να λαμβάνουν αποφάσεις βασιζόμενοι σε συγκεκριμένα δεδομένα και σε πλάνα δράσης που θα έχουν προκύψει από την ανάλυσή τους. Με τα μοντέλα πρόβλεψης, τα οποία δίνουν τη δυνατότητα δημιουργίας σεναρίων, θα ξέρουν τα αποτελέσματα από το κάθε πλάνο δράσης και θα μπορούν να επιλέγουν με περισσότερη ασφάλεια ποια στρατηγική θα πρέπει να ακολουθήσουν για να πετύχουν συγκεκριμένους στόχους. Επιπλέον, μέσα από την ανάλυση και ερμηνεία των Μεγάλων Δεδομένων θα έχουν τη δυνατότητα να ανακαλύψουν νέες ευκαιρίες και ανταγωνιστικά πλεονεκτήματα. Σύμφωνα με μία έρευνα της KPMG σε στελέχη επιχειρήσεων, το 86% αυτών απάντησαν ότι τα αναλυτικά μοντέλα τους βοήθησαν στην πιο γρήγορη λήψη αποφάσεων, το 67% ότι συνέβαλλαν να μειώσουν τον επιχειρηματικό κίνδυνο και το 80% είπε ότι παίρνουν πιο ακριβείς αποφάσεις [15].

3.2.4. Εξόρυξη δεδομένων στα κοινωνικά δίκτυα

Τα μέσα κοινωνικής δικτύωσης και οι σχετικές με αυτά πηγές είναι προφανές ότι έχουν εξελιχθεί ως η πιο χαρακτηριστική πηγή δεδομένων μεγάλου όγκου. Το Facebook, το Twitter, το Instagram, το Linkedin, το YouTube, το Tumblr, το Flickr, τα Wordpress blogs και φυσικά όλες οι εφαρμογές της Google έχουν αναδειχθεί σε κολοσσούς πληροφορίας η οποία διαδίδεται ταχύτατα και βρίσκει χρησιμότητα σε διαφορετικές περιοχές και βιομηχανίες εφόσον συγκεντρωθεί και αναλυθεί με τον κατάλληλο τρόπο. Κάποια από τα πεδία που εμπλέκονται στην ανάλυση δεδομένων μεγάλου όγκου από τα μέσα κοινωνικής δικτύωσης και άλλες διαδικτυακές πηγές είναι η εξόρυξη δεδομένων (data mining), η μηχανική μάθηση (machine learning), η στατιστική (statistics), η εξόρυξη πληροφορίας από γράφους (graph mining), η άντληση πληροφορίας (information retrieval), η γλωσσολογία (linguistics), η επεξεργασία φυσικής γλώσσας (natural language processing – NLP), ο σημασιολογικός ιστός (semantic Web), οι οντολογίες (ontologies) και οι υπολογιστικές μέθοδοι για δεδομένα μεγάλου όγκου (big data computing) [16].

Οι τεχνικές εξόρυξης δεδομένων που έχουν αναπτυχθεί, έχει αποδειχθεί ότι μπορούν πολύ αποδοτικά να διαχειριστούν τα ιδιαίτερα χαρακτηριστικά των δεδομένων μεγάλου όγκου που κατακλύζουν τα μέσα κοινωνικής δικτύωσης (μέγεθος, θόρυβο και δυναμική αλλαγή). Η φύση των δεδομένων αυτών απαιτεί αυτοματοποιημένες διαδικασίες ανάλυσης οι οποίες πραγματοποιούνται σε πολύ μικρό χρόνο. Ενδιαφέρον παρουσιάζει το γεγονός ότι όπως τα μέσα κοινωνικής δικτύωσης απαιτούν αυτοματοποιημένους τρόπους επεξεργασίας της πληροφορίας λόγω του τεράστιου όγκου των δεδομένων τους, αντίστοιχα οι τεχνικές που έχουν αναπτυχθεί απαιτούν αξιοσημείωτο όγκο δεδομένων για να εντοπίσουν μοτίβα και να παράγουν αξιόπιστα αποτελέσματα. Τα δεδομένα μεγάλου όγκου των κοινωνικών δικτύων, εν ολίγοις, αποτελούν άριστο πεδίο εφαρμογής για της σύγχρονες τεχνικές εξόρυξης δεδομένων [17].



Εικόνα 5: Ο εννοιολογικός χάρτης των Big Data των κοινωνικών δικτύων [18]

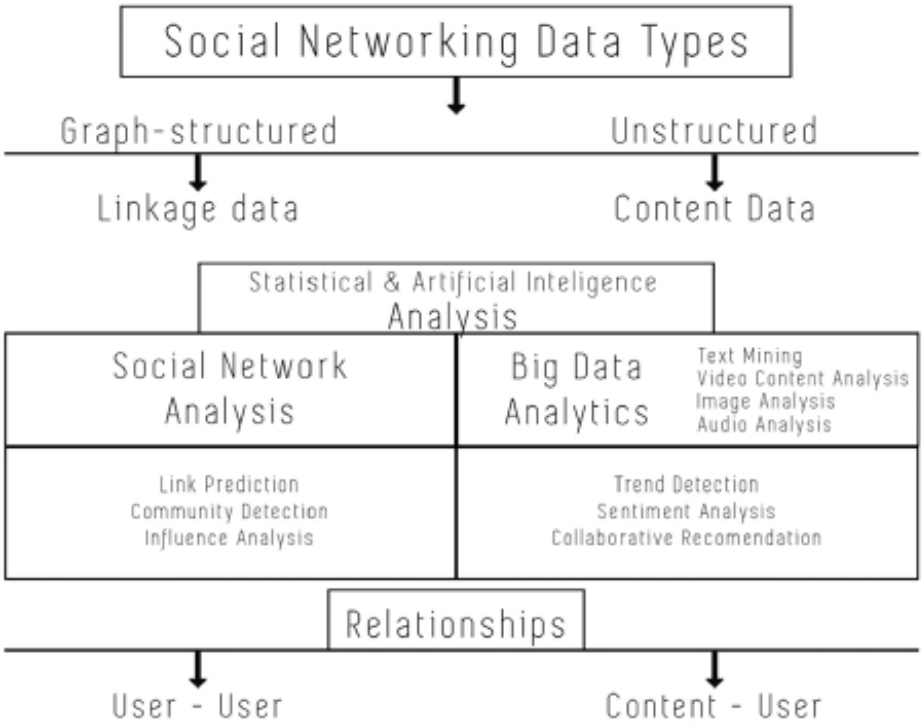
Οι πληροφορίες που συναντάμε στα μέσα κοινωνικής δικτύωσης ανήκουν στην κατηγορία των **δομημένων** (structured) ή **αδόμητων** (unstructured) δεδομένων, ανάλογα με το αν είναι οργανωμένα με έναν προκαθορισμένο τρόπο ή όχι. Τα δεδομένα που θεωρούνται δομημένα είναι τοποθετημένα σε ένα αρχείο με σταθερά πεδία ή μεταβλητές, και υπάρχει η δυνατότητα να αποθηκεύονται σε βάσεις δεδομένων με αρκετά μεγάλη χωρητικότητα, ευελιξία και συνέπεια (συνήθως MySQL, Microsoft SQL, Oracle).

Εκτός από τα δομημένα δεδομένα, υπάρχουν και τα αδόμητα δεδομένα, τα οποία δεν είναι εύκολο για ένα πρόγραμμα υπολογιστή να τα διαβάσει, να τα επεξεργαστεί, να τα ταξινομήσει, να τους αλλάξει σειρά, να μετρήσει τις τιμές και να προσθέσει περισσότερες παρατηρήσεις (πχ. απλό αρχείο κειμένου). Τέλος υπάρχουν και ημι-δομημένα δεδομένα, τα οποία δεν είναι μεν σε σταθερά πεδία, ωστόσο τα πεδία αυτά ξεχωρίζουν και τα δεδομένα εξακολουθούν να είναι αναγνωρίσιμα. Δύο κοινές μορφές είναι τα δεδομένα σε XML (Extensive Markup Language) ή JSON (JavaScript Object Notation) μορφές. Τα ημι-δομημένα και αδόμητα δεδομένα συνήθως λειτουργούν με NoSQL βάσεις δεδομένων (και όχι μόνο με SQL). Ο λόγος που χρησιμοποιούνται αυτές οι βάσεις δεδομένων είναι επειδή είναι εξαιρετικά ευέλικτες και μπορούν να χειριστούν ένα ευρύ φάσμα μορφών δεδομένων.

Στα κοινωνικά μέσα, οι σχέσεις μεταξύ των χρηστών είναι δομημένα δεδομένα ενώ το περιεχόμενο των αναρτήσεων δεν είναι. Συνήθως, τα δομημένα δεδομένα των κοινωνικών δικτύων μπορούν να

μοντελοποιηθούν με ένα γράφο G=(V,E) όπου V το σύνολο των κόμβων του δικτύου (φυσικά πρόσωπα, οργανισμοί, προϊόντα, κ.ο.κ) και E το σύνολο των ακμών που συνδέει τους κόμβους μεταξύ τους και αναπαριστά τις σχέσεις τους. Εφαρμόζοντας τεχνικές ανάλυσης γράφων μπορούμε να αντλήσουμε χρήσιμη πληροφορία από τα διασυνδεδεμένα δομημένα δεδομένα.

Αντίστοιχα, το περιεχόμενο που δημιουργούν οι χρήστες (user generated content) και συμπεριλαμβάνει αναρτήσεις, tweets, likes, αναρτήσεις πολυμέσων (multimedia posts), ανήκει στα αδόμητα δεδομένα και χρίζει ανάλυσης προσανατολισμένης στο περιεχόμενο (content-based analysis). Οι επιστημονικοί κλάδοι που κυρίως εμπλέκονται στην τελευταία είναι η Τεχνητή Νοημοσύνη και η Στατιστική [19]. Παρόλο που τα τελευταία χρόνια έχουν γίνει σημαντικά βήματα, η ανάλυση των δεδομένων από τα μέσα κοινωνικής δικτύωσης από αδόμητες ή ημι-δομημένες πηγές προκειμένου να εξαχθεί χρήσιμη πληροφορία δεν έχει ακόμη επιλυθεί πλήρως σαν πρόβλημα. Οι κλασικές μέθοδοι, τα εργαλεία και οι πλατφόρμες για τη διαχείριση των δεδομένων δεν επαρκούν, καθώς δημιουργούνται προβλήματα σχετικά με διαφορετικές πτυχές του θέματος όπως η αναπαράσταση της γνώσης, η συλλογή, η ομογενοποίηση, η επεξεργασία, η ανάλυση και η οπτικοποίηση του τεράστιου όγκου δεδομένων [20].



Εικόνα 6: Τύποι δεδομένων και ανάλυσης στα κοινωνικά δίκτυα [18]

Πίνακας 1: Τεχνικές Εξόρυξης Δεδομένων στα κοινωνικά μέσα [18]

Τεχνική	**Πλεονεκτήματα**	**Μειονεκτήματα**
SVM (Support Vector Machine)	Μία από τις καλύτερες τεχνικές για την επίλυση προβλημάτων κατηγοριοποίησης (classification problems). Έχει καλή απόδοση με χρήση πολλών χαρακτηριστικών (features) και μικρό όγκο δεδομένων εκπαίδευσης (training dataset). Είναι κατάλληλη για ομαδοποίηση εκτός σύνδεσης (offline clustering).	Έχει το πρόβλημα των αραιών συνδέσμων περιεχομένου (sparse context links).
ANN (Artificial Neural Network)	Self-Organizing Map (SOM): Δυνατότητες υψηλού επιπέδου που διευκολύνουν την ανάλυση δεδομένων πολλών διαστάσεων (high-dimensional data analysis). Έχει οπτικά πλεονεκτήματα.	Μέσος SOM: Επιφέρει χάρτες χαμηλότερης ποιότητας από αυτούς που αποκτώνται από την έκδοση πυρήνα..
DT (Decision Trees)	Random Forest (RF): Αποτελεσματικό στις εκτιμήσεις σχετικά με το ποιές μεταβλητές είναι σημαντικές στην ταξινόμηση. Ισχυρή τεχνική η οποία αποδίδει καλά σε πληθώρα διαδικασιών μηχανικής μάθησης (learning tasks).	
BN (Bayesian Networks)	Πολύ αποτελεσματικά για ομαδοποίηση κειμένου. Απλός αλγόριθμος ταξινόμησης. Πολύ αποδοτική μέθοδος από την άποψη της υπολογιστικής πολυπλοκότητας.	
k-NN (k-nearest Neighbors)	Μία από τις πιο απλές και ξεκάθαρες τεχνικές ταξινόμησης στην αναγνώριση μοτίβων.	Χαμηλότερη απόδοση σε μικρά σύνολα δεδομένων. Η απόδοση φθίνει για δεδομένα με πολλά χαρακτηριστικά. Υπάρχει εξάρτηση από το επιλεγμένο χαρακτηριστικό και την απόσταση (feature and distance measure).
Fuzzy	Εξειδικεύεται στην μοντελοποίηση αόριστων φάσεων κοινωνικής λογικής και λαμβάνει υπόψιν τον στοχαστικό παράγοντα της ανθρώπινης λογικής.	Απαιτεί εξειδίκευμένη γνώση πάνω στο σημασιολογικό ιστό και την ασαφή λογική (semantic web and fuzzy systems) για να είναι εφικτός ο χειρισμός του semantic fuzzy rule μέσα από μια διαδικασία εκτός σύνδεσης.

K-Means	k-Medoids: Λιγότερο ευαίσθητη σε ακραίες τιμές. Χρησιμοποιεί όσο λιγότερες ομάδες γίνεται και αποτυπώνει στατιστικά και εμπορικά σημαντικά χαρακτηριστικά των ομάδων. Κατάλληλη για καθορισμένο αριθμό ομάδων με άγνωστα χαρακτηριστικά, με βάση μεταβλητές που μπορούν να οριστούν. SK-Means: Αποδοτική από άποψη χρόνου. Αποδοτική με σύνολα δεδομένων πολλών διαστάσεων (high-dimensional datasets). Μπορεί να εκτελεσθεί παράλληλα με αποδοτικό τρόπο και συγκλίνει γρήγορα σε τοπικό μέγιστο. Μπορεί να αποτελέσει ένα μοντέλο για επαναχρησιμοποίηση σε μελλοντικές ταξινομήσεις.	k-Medoids: Απαιτεί τον αριθμό των ομάδων σαν όρισμα. Όσο μεγαλώνει ο αριθμός των ομάδων, η ποιότητά τους φθίνει. Συχνά συγκλίνει σε τοπικό ελάχιστο.
DBA (Density Based Algorithm)	DBSCAN: Δεν απαιτεί προκαθορισμένο αριθμό ομάδων και φίλτρο θορύβου. Αντιμετωπίζει το θόρυβο σαν ακραίες τιμές που δε μπορούν να ενταχθούν σε καμία ομάδα (cluster). Ικανότητα ανίχνευσης ομάδων αυθαίρετου σχήματος (arbitrary-shaped clusters)	DBSCAN: Περιλαμβάνει όλα τα density-reachable σημεία σε μία ομάδα Ακατάλληλη για κάποιες εφαρμογές γιατί δεν υπάρχει υπόθεση για τον αριθμό των ομάδων με ορισμένα θέματα.
LDA (Linear Discriminant analysis)	Επιπλέον της ομαδοποίησης των δεδομένων, χαρακτηρίζει τα κείμενα. Χρήσιμη για την ανάπτυξη εφαρμογών πολυμέσων. Σχεδιασμένη για να εκμεταλλεύεται τη συχνότητα των όρων.	Συχνά συγκλίνει σε τοπικό ελάχιστο. Αντιμετωπίζει το πρόβλημα των αραιών συνδέσμων.
Wrapper		Απαιτεί στρατηγικές με υψηλό επίπεδο αυτοματοποίησης. Η συντήρηση του wrapper γίνεται δύσκολη καθώς μεγαλώνει το Διαδίκτυο.
HC (Hierarchical clustering)		Δεν προσαρμόζεται στην αύξηση των δεδομένων, διότι στηρίζεται σε έναν πλήρως προσδιορισμένο πίνακα ομοιοτήτων.

Σύμφωνα με την έρευνα των MohammadNoor Injadat, Fadi Salo, Ali Bou Nassif [18] σε 66 επιλεγμένα άρθρα, οι SVM, BN και DT είναι οι επικρατέστερες τεχνικές εξόρυξης δεδομένων που εφαρμόζονται στα μέσα κοινωνικής δικτύωσης με ποσοστό χρησιμοποίησης 51%. Επίσης, τα μέσα κοινωνικής δικτύωσης και ο τομέας των επιχειρήσεων (business and management) είναι οι πιο ενεργοί κλάδοι στους οποίους βρίσκει εφαρμογή η εξόρυξη δεδομένων, σε ποσοστό 79%.

3.3. Διαδικασία Ανίχνευσης Τάσης

Στα πλαίσια της Επιχειρηματικής Ευφυΐας, και συγκεκριμένα στον εντοπισμό των ευκαιριών και τη λήψη αποφάσεων σχετικά με αυτές, μπορούμε να εντάξουμε το ζήτημα της ανίχνευσης των τάσεων (trends) της αγοράς στον κλάδο της επιχείρησης. Η παραδοσιακή μέθοδος για την ανίχνευση των τάσεων μπορεί να θεωρηθεί ότι διαχρονικά είναι η Έρευνα Αγοράς, όπου ένα τυχαίο ή μη δείγμα καταναλωτών εκφράζει τις προτιμήσεις του απαντώντας ευθέως σε ερωτήσεις. Σε αυτή τη μέθοδο υπάρχει ένα σημαντικό μειονέκτημα, καθώς είναι πολύ πιθανό ότι το υπόβαθρο και η συμπεριφορά του δείγματος τη στιγμή εκείνη εμπλέκονται στα αποτελέσματα της έρευνας, εισάγοντας σε αυτά μια προκατάληψη [21]. Όπως έχει προαναφερθεί, η τεχνολογία σήμερα επιτρέπει την πρόσβαση στον τεράστιο όγκο πληροφοριών που υπάρχουν στο διαδίκτυο, όπου ένας σημαντικός αριθμός ατόμων εκφράζουν ελεύθερα τη γνώμη τους [22]. Επομένως το πρόβλημα της παραδοσιακής μεθόδου παρακάμπτεται, και οι επιχειρήσεις έχουν μία ανεξάντλητη πηγή πληροφοριών για να χρησιμοποιήσουν σε όλους τους τομείς της Επιχειρηματικής Ευφυΐας, συμπεριλαμβανομένης της ανίχνευσης των τάσεων.

3.3.1. Ορισμοί

Για να διερευνήσουμε τις μεθόδους ανίχνευσης των τάσεων της αγοράς, θα πρέπει να τις εντάξουμε σε μια ευρύτερη κατηγορία ώστε να δούμε τι τεχνολογική πρόοδος έχει γίνει στον κλάδο γενικά, και ποιες μεθοδολογίες και εργαλεία μπορούν να παραμετροποιηθούν κατάλληλα ώστε να εξυπηρετήσουν τους σκοπούς της συγκεκριμένης διπλωματικής. Αν και η έννοια της τάσης είναι αρκετά ευρεία και μια τάση είναι από μόνη της εκ φύσεως μεταβαλλόμενη, είναι απαραίτητο να ορισθεί με αυστηρό τρόπο, έτσι ώστε να μοντελοποιηθεί για να είναι ανιχνεύσιμη από εξειδικευμένες μεθόδους μέσα από το μεγάλο όγκο των δεδομένων του διαδικτύου (big data). Συνήθως σαν τάση αναφέρεται κάποιο σημαντικό γεγονός ή κάποιο δημοφιλές θέμα, και η δημοφιλία είναι μετρήσιμη με διαφορετικά μέτρα ανάλογα με την εκάστοτε έρευνα ή μεθοδολογία. Μερικά παραδείγματα ορισμών δίνονται παρακάτω:

Τάση (Ορισμός 1)

Ως τάση ορίζεται ένα δοθέν γεγονός ή θέμα του οποίου η επίδραση σε ένα σύστημα ως όλον, είναι πάνω από το μέσο όρο για μια συγκεκριμένη χρονική περίοδο. Επιπρόσθετα, για ένα σύστημα αποτελούμενο από μια αλυσίδα γεγονότων, η τάση προσδιορίζεται από την αναμενόμενη μελλοντική συμπεριφορά δεδομένης της συμπεριφοράς που είχε στο παρελθόν και της ανταπόκρισής του στα εξωτερικά ερεθίσματα.[23]

Τάση (Ορισμός 2)

Ως τάση ή δημοφιλές θέμα στα μέσα κοινωνικής δικτύωσης μπορεί να οριστεί το θέμα το οποίο παρουσιάζει απότομη και σημαντική αύξηση στον αριθμό των αναμεταδόσεών του. [24]

Τάση (Ορισμός 3)

Ως σχετική τάση στα μέσα κοινωνικής δικτύωσης μπορεί να οριστεί το θέμα το οποίο μπορεί να μην είναι δημοφιλές σύμφωνα με τον απόλυτο αριθμό αναμεταδόσεών του σε σχέση με άλλα θέματα, αλλά παρουσιάζει απότομη και σημαντική αύξηση στον αριθμό των αναμεταδόσεών του σε σχέση με το αναμενόμενο.[24] Αυτή η οπτική μπορεί να είναι χρήσιμη για τον έγκαιρο εντοπισμό ή την πρόβλεψη των τάσεων, καθώς ανιχνεύει τη δυναμική ενός θέματος ανεξάρτητα από το πόσο έχει διαδοθεί μέχρι στιγμής.

Τάση (Ορισμός 4)

Ως τάση ή ανερχόμενο θέμα (emerging topic) ορίζεται ως η θεματική περιοχή η οποία παρουσιάζει ανοδική πορεία σε ενδιαφέρον (interest) και χρησιμότητα (utility) μέσα στο χρόνο. [25]

Πρόβλημα Ανίχνευσης Τάσεων (Ορισμός 1)

Δεδομένου ενός συνόλου θεμάτων, ζητείται να προσδιορισθεί αν ο τρόπος που συμπεριφέρονται στη διάρκεια του χρόνου τα καθιστά τάση. [23]

Στο μεγαλύτερο μέρος της βιβλιογραφίας σχετικά με την ανίχνευση τάσεων στο διαδίκτυο, ο όρος «τάση» (trend) περιγράφεται επιτυχώς από τους παραπάνω ορισμούς, καθώς συνήθως αναφέρονται ως τάσεις τα ανερχόμενα/δημοφιλή θέματα (trending/emerging/hot topics), δηλαδή τα θέματα που εμφανίζονται ή πρόκειται να συζητηθούν περισσότερο. Στα πλαίσια της παρούσας εργασίας θα ασχοληθούμε με τις τάσεις της αγοράς σε κάποιο **συγκεκριμένο κλάδο**, όπου ο ορισμός είναι και πάλι ακριβής, λαμβάνοντας όμως υπόψιν ότι τα δεδομένα από τα οποία θα πρέπει να εξαχθούν συμπεράσματα σχετικά με τις τάσεις είναι ήδη επιλεγμένα ώστε να αφορούν μια συγκεκριμένη θεματολογία. Θα αναζητήσουμε δηλαδή ***μεθόδους ανίχνευσης των τάσεων*** αλλά που αφορούν ένα συγκεκριμένο θέμα (topic), άμεσα συνυφασμένο με τον εκάστοτε κλάδο.

Πρόβλημα Ανίχνευσης Τάσεων (Ορισμός 2)

Δοσμένης μιας χρονικά ταξινομημένης ροής αναρτήσεων των χρηστών Pt, t = [1,… ∞), η οποία φτάνει σε πραγματικό χρόνο (tweets) η με μια συγκεκριμένη χρονική συχνότητα (blog posts), προσδιόρισε τα θέματα και τις σχετικές αναρτήσεις που είναι δημοφιλή (trending) σε κάθε χρονική στιγμή, και παρακολούθησε την εξέλιξή τους μέσα στο χρόνο στο θέμα της δημοτικότητάς τους. [26]

Θέμα (topic)

Ένα σύνολο σημαντικών φράσεων που ομαδοποιούνται μαζί με βάση την ομοιότητά τους. [27]

3.3.2. Πηγές

Τα δεδομένα των συστημάτων Επιχειρηματικής Ευφυΐας μπορεί να περιλαμβάνουν ιστορικά στοιχεία (αρχείου) ή νέα δεδομένα που συγκεντρώνονται ακριβώς για τους σκοπούς της ανάλυσης σε σύντομο χρονικό διάστημα από τη στιγμή που δημιουργούνται, βοηθώντας την ανάλυση να υποστηρίξει τη διαδικασία λήψης αποφάσεων σε στρατηγικό και πρακτικό επίπεδο.

Οι διαδικτυακές πηγές απ’ όπου μπορεί να εξαχθεί χρήσιμο περιεχόμενο σχετικά με τις προτιμήσεις των καταναλωτών και με τις τρέχουσες τάσεις σε ένα θέμα ή σε έναν κλάδο, είναι όλες οι εφαρμογές του Web 2.0 όπου οι χρήστες εκφράζονται ελεύθερα δημιουργώντας περιεχόμενο στον ιστό (user generated content). Αυτές μπορεί να είναι ιστολόγια (blogs) χρηστών, forums, κριτικές (reviews)

προϊόντων και υπηρεσιών σε ηλεκτρονικά καταστήματα και αντίστοιχες πλατφόρμες, και τέλος τα microblogs, ένας όρος που χρησιμοποιείται για να εκφράσει τις μικρού μήκους αναρτήσεις των χρηστών στα μέσα κοινωνικής δικτύωσης. Ανάλογα με τον εκάστοτε τομέα της αγοράς αλλά και το θέμα ή την κατηγορία προϊόντος/υπηρεσίας, διαφορετικές πηγές έχουν διαφορετική βαρύτητα.

Ωστόσο, τα τελευταία χρόνια το ενδιαφέρον του ακαδημαϊκού αλλά και επιχειρηματικού κόσμου έχει επικεντρωθεί στην ανάλυση των δεδομένων από τα μέσα κοινωνικής δικτύωσης ως κύρια πηγή περιεχομένου δημιουργημένο από χρήστες. Αυτό μεταξύ άλλων οφείλεται στην ευρεία και ολοένα αυξανόμενη χρήση τους ως μέσο έκφρασης απόψεων, προτιμήσεων και συναισθημάτων, ως μέσο αλληλεπίδρασης και άσκησης επιρροής μεταξύ των χρηστών αλλά και ως μέσο προώθησης και στοχευμένης διαφήμισης από την πλευρά των εταιρειών. Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης όπως έχουν διαμορφωθεί αποτελούν ένα ολοκληρωμένο περιβάλλον όπου οι επιχειρήσεις μπορούν να παρακολουθήσουν την ανταπόκριση του κοινού στα προϊόντα και τις υπηρεσίες τους, να σχεδιάσουν και να οργανώσουν τις διαφημιστικές τους καμπάνιες, να διαχειριστούν την εταιρική τους εικόνα και τις δημόσιες σχέσεις τους αλλά και να έχουν εικόνα της αγοράς και του ανταγωνισμού σε πραγματικό χρόνο [27]. Σε αυτά τα πλαίσια, είναι αναμενόμενο ότι και η πορεία των τάσεων της αγοράς που διαμορφώνονται (από τους χρήστες αλλά και τις εταιρείες) και τελικά υιοθετούνται ή απορρίπτονται από το ευρύ κοινό, απεικονίζεται με ακρίβεια στα μέσα κοινωνικής δικτύωσης για τους περισσότερους κλάδους.

3.3.3. Ανίχνευση και πρόβλεψη τάσεων

Από τη σκοπιά της Επιχειρηματικής Ευφυΐας, οι τάσεις είναι χρήσιμο να ανιχνεύονται σε πρώιμο στάδιο, έτσι ώστε να υπάρχει χρόνος να αξιοποιηθούν προς όφελος της επιχείρησης. Στην ιδανική περίπτωση, τα διαθέσιμα μέσα μπορούν να χρησιμοποιηθούν για την **πρόβλεψη των τάσεων** (trend prediction), πριν ακόμα γίνουν γνωστές και υιοθετηθούν από το ευρύ κοινό ή/και τους ανταγωνιστές. Εάν αυτό δεν είναι εφικτό ή εάν ενέχει υψηλό ρίσκο, οι τάσεις είναι χρήσιμο να εντοπίζονται (trend detection) αρκετά έγκαιρα, δηλαδή ανάμεσα στο ξεκίνημα και την περίοδο μέγιστης απήχησης, και όχι ετεροχρονισμένα.

Ένας τρόπος προσέγγισης του ζητήματος της πρόβλεψης των τάσεων στα μέσα κοινωνικής δικτύωσης και όχι απλά της ανίχνευσής τους έχει προταθεί από τους D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto [28]. Στην δημοσίευση παρουσιάζεται μια μεθοδολογία για την εύρεση ατόμων με επιρροή στη δημιουργία τάσεων σε ένα συγκεκριμένο κλάδο (trendsetters). Πρόκειται για τα άτομα τα οποία υιοθετούν και διαδίδουν μία ιδέα/τάση πριν αυτή γίνει στην πράξη δημοφιλής, εν τη γενέσει της. Αντίθετα, τα άτομα τα οποία υιοθετούν την τάση καθυστερημένα αλλά την διαδίδουν ακόμα περισσότερο και επηρεάζουν ένα μεγάλο αριθμό χρηστών, ορίζονται σαν άτομα με επιρροή στη διάδοση των τάσεων (influencers). Επομένως ο εντοπισμός των trendsetters σχετίζεται άμεσα με την πρόβλεψη των τάσεων (trend prediction ή early trend detection) ενώ ο εντοπισμός των influencers σχετίζεται άμεσα με την ανίχνευση των τάσεων (trend detection).

Οι Yi Han , Binxing Fang, Yan Jia [29] αναπτύσσουν επίσης μια μέθοδο πρόβλεψης τάσεων στα μέσα κοινωνικής δικτύωσης λαμβάνοντας υπόψιν μεταξύ άλλων και τον γεωγραφικό παράγοντα. Η μέθοδός τους βασίζεται σε αρχές μηχανικής μάθησης και μπορεί να εφαρμοστεί σε δύο περιπτώσεις: α) Την παρακολούθηση της διάδοσης ενός θέματος μέσα στο μέσο (spreading trace), η οποία μπορεί να αναπαρασταθεί σαν ένα σύνολο κόμβων, λέξεων-κλειδιών και χρονοσφραγίδων. Το σύστημα σε αυτή την περίπτωση προβλέπει τη μελλοντική κατάσταση του δεδομένου θέματος. Και β) Την

πρόβλεψη θέματος που θα έχει μεγάλη επιρροή. Το σύστημα αυτόματα υπολογίζει την εξέλιξη μιας υποψήφιας χρονοσειράς, και έχει σαν έξοδο θέματα που πιθανώς θα έχουν επιρροή στο μέλλον.

Άλλη μια μέθοδος με στόχο την ανίχνευση ανερχόμενων θεμάτων προτείνεται από τους Qi Dang, Feng Gao, Yadong Zhou [30] Η μέθοδος αφορά συγκεκριμένα την έγκαιρη ανίχνευση ανερχόμενων θεμάτων (early detection of emerging topics) σε δίκτυα σύντομων αναρτήσεων (micro-blogging networks), και βασίζεται σε δίκτυα DBN (Dynamic Bayesian Networks). Το DBN μοντέλο υπολογίζει την πιθανότητα να είναι ανερχόμενη μια λέξη-κλειδί (keyword) και στη συνέχεια με βάση την συσχέτιση μεταξύ λέξεων-κλειδιών, εφαρμόζεται ο DBSCAN αλγόριθμος για να ομαδοποιηθούν οι ανερχόμενες λέξεις-κλειδιά σε ανερχόμενα θέματα. Για τα εν λόγω θέματα, στην παρούσα δημοσίευση εντοπίζονται δύο χαρακτηριστικά: α) ελκυστικότητα (attractiveness), η οποία αποτελεί το βαθμό στον οποίο το θέμα έλκει τους χρήστες να το διαδώσουν και β) κόμβους-κλειδιά (key-node), χαρακτηριστικό που αφορά την ύπαρξη χρηστών με επιρροή οι οποίοι μπορούν να προκαλέσουν μεγάλο αριθμό αναμεταδόσεων και να διαδώσουν σημαντικά το θέμα. Αξίζει να αναφερθεί πως στα περισσότερα παραδείγματα που δίνουν οι ερευνητές αναφέρονται σε θέματα-γεγονότα και όχι τάσεις, καθώς και σε θέματα που παίρνουν πολύ γρήγορα από την εμφάνισή τους μεγάλες διαστάσεις, σε αντίθεση με την αμέσως προηγούμενη έρευνα όπου ο χρόνος ωρίμανσης του ανερχόμενου θέματος μπορεί να είναι και αρκετές ημέρες.

3.3.4. Στάδια

Μετά την επισκόπηση της σχετικής βιβλιογραφίας, μπορεί να παρατηρηθεί ότι οι περισσότερες μεθοδολογίες και τα εργαλεία που στοχεύουν στην Ανίχνευση Τάσεων (trend detection) αναφέρονται στις εξής **κατηγορίες**:

1. **Εντοπισμό Ανερχόμενων Θεμάτων** ([30], [31], [17], [26], [32], [27], [33], [34], [35], [36], [37])
2. **Εντοπισμό Σημαντικών Γεγονότων** ([23], [38], [39], [40], [41], [42])
3. **Ανίχνευση Τάσεων για προκαθορισμένα θέματα** ([24] [43], [29])

Συνήθως υπάρχει μια εξειδίκευση ανά μεθοδολογία ή εργαλείο, με βάση την οποία γίνεται και η επιλογή των πηγών, των αλγορίθμων και η εφαρμογή των σταδίων ανά περίπτωση. Κάποια παραδείγματα είναι η έγκαιρη ανίχνευση θεμάτων ή γεγονότων (early trend detection ή trend prediction), εννοώντας τη προσπάθεια εντοπισμού της τάσης σε αρχικό στάδιο διάδοσής της (π.χ. [30], [34], [29], [42]) , ο θεματικός προσδιορισμός (πχ. ανερχόμενα πολιτικά ζητήματα, [33]) ή κάποιος άλλος προσδιορισμός (πχ. ανίχνευση ανερχόμενου θέματος ανά γεωγραφική περιοχή, [35]).

Παρά τις διαφορές και την ποκιλία που υπάρχει ανάλογα με τη σκοπιμότητα της μεθόδου και το είδος της τάσης που επιδιώκεται να ανιχνευθεί, μπορούμε να προσδιορίσουμε κάποια γενικά στάδια τα οποία είναι κοινά σε πολλές από τις μεθοδολογίες και λογικά αναμενόμενα. Σαν **στάδια** επομένως της Ανίχνευσης Τάσεων σε μέσα κοινωνικής δικτύωσης και διαδικτυακές πηγές αναφέρουμε:

1. **Συλλογή δεδομένων** (Data Collection)
2. **Επεξεργασία ή προ-επεξεργασία δεδομένων** (Data preprocessing)
3. **Ανίχνευση θέματος** (Topic Detection)
4. **Προσδιορισμός τάσης** (Trend Detection)

5. **Ανάλυση συναισθήματος** (Sentiment Analysis)

Από αυτά η συλλογή δεδομένων και η επεξεργασία τους έχουν τη μεγαλύτερη συχνότητα εμφάνισης, ενώ η ανίχνευση θέματος και ο προσδιορισμός τάσης μπορεί να εναλάσσονται, ή ακόμα και να συνυπάρχουν στο ίδιο βήμα της μεθόδου. Τέλος, η ανάλυση συναισθήματος είναι το στάδιο που εμφανίζεται λιγότερο, καθώς χρησιμοποιείται σε συγεκριμένες περιπτώσεις. Στα επόμενα υποκεφάλαια ακολουθεί ανάλυση του καθενός από τα προαναφερθέντα στάδια και των τεχνικών που χρησιμοποιούνται κατά την υλοποίησή του.

3.3.4.1. Συλλογή δεδομένων

Η συλλογή δεδομένων από το διαδίκτυο (Web Data Extraction) είναι ένα σημαντικό πεδίο έρευνας το οποίο έχει προσεγγιστεί με διάφορα εργαλεία και βρίσκει εφαρμογή σε πολλαπλές περιπτώσεις. Για ευνόητους λόγους αποτελεί προαπαιτούμενο για την Εξαγωγή Γνώσης από τα δεδομένα, και κατ επέκταση την Εξόρυξη Δεδομένων και την Ανίχνευση Τάσεων. Αυτό διότι η συνηθέστερη περίπτωση είναι να μην υπάρχει έτοιμο σύνολο δεδομένων προς επεξεργασία και ανάλυση.

Τα συστήματα συλλογής ή εξαγωγής δεδομένων (Web Data Extraction systems) είναι μια ευρεία κατηγορία εφαρμογών λογισμικού οι οποίες στοχεύουν στην συγκέντρωση των δεδομένων από μία ή περισσότερες διαδικτυακές πηγές [44]. Το σύστημα αλληλεπιδρά με την πηγή και εξάγει δεδομένα από αυτή ανάλογα με τις εκάστοτε ανάγκες. Για παράδειγμα, αν πρόκειται για μια σελίδα HTML, το περιεχόμενο προς εξαγωγή μπορεί να είναι κάποια στοιχεία της σελίδας ή και ολόκληρο το κείμενο που περιέχεται σε αυτή. Στην περίπτωση των κοινωνικών δικτύων μπορεί να είναι το κείμενο των αναρτήσεων που περιέχουν κάποιες συγκεκριμένες λέξεις ή συγκεκριμένα hashtags, οι αναρτήσεις που πληρούν κάποια χαρακτηριστικά (π.χ. αριθμός αναμεταδόσεων, χρονική ή γεωγραφική σήμανση) κ.ο.κ.

Πέρα από τις μεθοδολογίες για ανίχνευση τάσεων οι οποίες βρίσκονται σε πρώιμο ή ερευνητικό στάδιο και εφαρμόζονται πάνω σε έτοιμα σύνολα δεδομένων, οι περισσότερες μέθοδοι περιλαμβάνουν και το στάδιο της συγκέντρωσης των δεδομένων.

Διακρίνουμε τις εξής βασικές περιπτώσεις **τεχνικών συλλογής δεδομένων**:

1. Με τη βοήθεια κάποιου **Προγράμματος Ανίχνευσης Ιστού** ([23], [26], [27], [33], [34],[36], [45])
2. Μη τη χρήση ερωτήματος αναζήτησης ή απλού αλγορίθμου **απευθείας στις Διεπαφές Προγραμματισμού των κοινωνικών μέσων** ([42], [41], [40], [35], [31]) .

Αφού συλλεχθούν τα δεδομένα επιλέγεται είτε να προχωρήσει η μεθοδολογία στο αμέσως επόμενο στάδιο είτε να αποθηκευτούν σε κάποια βαση δεδομένων. Στην περίπτωση που η μεθοδολογία αφορά δεδομένα από παραπάνω από μία πηγές ή σε παραπάνω από μία μορφές, μπορεί να μεσολαβήσει επίσης ένα στάδιο ομογενοποίησης των δεδομένων και να αποθηκευτούν στη βάση σε ενιαία μορφή για να είναι ευκολότερη η επεξεργασία τους.

Προγράμματα Ανίχνευσης Ιστού

Τα Προγράμματα Ανίχνευσης Ιστού (Web Crawlers, Web Spiders, Web Parsers) έχουν αποδειχτεί αποδοτικά εδώ και πολλά χρόνια στην ανάλυση περιεχομένου του διαδικτύου. Ένα Πρόγραμμα

Ανίχνευσης (γνωστό και ως ρομπότ διαδικτύου) είναι ένα σενάριο (script) το οποίο σκανάρει τον ιστό με αυτοματοποιημένο και μεθοδικό τρόπο, και χρησιμοποιείται από ιστοσελίδες και μηχανές αναζήτησης για να ανασύρει αποτελέσματα. Συνήθως είναι αρκετά περίπλοκο στο σχεδιασμό και την υλοποίηση. Στην περίπτωση των κοινωνικών δικτύων, ένας τρόπος με τον οποίο μπορεί να λειτουργήσει ένα τέτοιο πρόγραμμα είναι να αναζητήσει τις σχέσεις μεταξύ των χρηστών – κόμβων του δικτύου και να δημιουργήσει μια λίστα την οποία θα ακολουθήσει. Ο λογαριασμός χρήστη από τον οποίο ξεκινά η προσπέλαση αποτελεί τον αρχικό κόμβο της αναζήτησης (seed node) και ο αλγόριθμος σταματά είτε όταν προσπελαστεί όλο το δέντρο είτε όταν γίνει αληθής κάποια προκαθορισμένη συνθήκη.

Η διαδικασία αυτή σε όρους θεωρίας γράφων, αντιστοιχεί σε προσπέλαση του δικτύου με αναζήτηση κατά βάθος (Breadth-First-Search, BFS), και παράξει αξιόπιστα αποτελέσματα για κοινωνικούς γράφους που μοντελοποιούνται χωρίς βάρη (unweighted social graphs). Για το λόγο αυτό έχει χρησιμοποιηθεί σε πολλές μελέτες που αφορούν την τοπολογία των κοινωνικών δικτύων ([46], [47], [48], [49], [50]).

Διεπαφές Προγραμματισμού των δικτύων (API)

Σχετικά με την δεύτερη προσέγγιση, ξεκινάμε από το γεγονός ότι οι πλατφόρμες των κοινωνικών ιστών διαθέτουν σύγχρονες **Διεπαφές Προγραμματισμού/API** (π.χ. Twitter API, Facebook API, Sina API) πολλές φορές διαθέσιμες σε παραπάνω από μία γλώσσες προγραμματισμού, οι οποίες επιτρέπουν την άμεση και εύκολη εξαγωγή δεδομένων από την ίδια την πλατφόρμα.

Τα δεδομένα αυτά περιλαμβάνουν τόσο τη δομή του εκάστοτε γράφου, δηλαδή τις σχέσεις μεταξύ των χρηστών όσο και το περιεχόμενο που παράγουν οι χρήστες. Για παράδειγμα το **Twitter API** δίνει πρόσβαση σε ολόκληρο το γράφο του δικτύου, επιτρέποντας στον ενδιαφερόμενο χρήστη ή σε κατάλληλο αλγόριθμο μέχρι και 150 ερωτήματα την ώρα στη βάση του [44]. Για ακόμα πιο υψηλές απαιτήσεις, χρήστες εγγεγραμένοι σε ειδική λίστα (white lists) μπορούν να παραγματοποιήσουν μέχρι και 20.000 ερωτήματα ανά διεύθυνση δικτύου (IP) την ώρα. Οι αυξημένες αυτές προγραμματιστικές δυνατότητες επιτρέπουν τη δημιουργία συνόλων δεδομένων (dataset) που περιλαμβάνουν δεκάδες εκατομμύρια προφίλ χρηστών του δικτύου και εκατοντάδες εκατομμύρια αναρτήσεις και επιτρέπουν την παρακολούθηση του δικτύου σε πραγματικό χρόνο [51].

Από την άλλη το **Facebook Graph API** παρέχει πρόσβαση στον «κοινωνικό γράφο» του δικτύου μέσω μιας αναπαράστασης των αντικειμένων που τον αποτελούν (χρήστες, αναρτήσεις, σελίδες, κ.ο.κ.). Οι αναρτήσεις που αποτελούν την κύρια πηγή δεδομένων για εξαγωγή πληροφορίας, περιέχουν σαν αντικείμενα του γράφου λεπτομέρειες του περιεχομένου της ανάρτησης, το προφίλ του χρήστη οπου έχει ανέβει, τον τύπο του μηνύματος όπως προδιορίζεται από το Facebook (φωτογραφία, κοινοποίηση κατάστασης, σύνδεσμος, κ.ο.κ.), την ημερομηνία δημιουργίας, τη συσκευή μέσω της οποίας δημοσιεύτηκε κ.α. Δεν προσφέρεται όμως, σε αντίθεση με το Twitter, η δυνατότητα ανάσυρσης των αναρτήσεων σε πραγματικό χρόνο μέσω συνεχούς ροής (real-time stream) οπότε επιβάλλεται η υλοποίηση ενός απλού αλγορίθμου αναζήτησης που θα ανασύρει τις αναρτήσεις ανά τακτά χρονικά διαστήματα [17].

3.3.4.2. Επεξεργασία δεδομένων

Σαν δεύτερο στάδιο της διαδικασίας για την αναγνώριση των τάσεων μπορεί να θεωρηθεί η **επεξεργασία** ή **προ-επεξεργασία** (preprocessing) των δεδομένων που έχουν συγκεντρωθεί. Το βήμα αυτό είναι απαραίτητο όπως και το προηγούμενο για κάθε είδους εξαγωγής πληροφορίας και Εξόρυξης Δεδομένων. Πρόκειται για τις ενέργειες που γίνονται ώστε να προετοιμαστούν τα δεδομένα για περαιτέρω ανάλυση. Αυτό μπορεί να σημαίνει τον «καθαρισμό» ή φιλτράρισμα των δεδομένων, και την επεξεργασία τους ώστε να αποκτήσουν μια μορφή πάνω στην οποία θα μπορούν να λειτουργήσουν αποδοτικά οι πιο εξειδικευμένοι αλγόριθμοι που θα ακολουθήσουν για την Ανίχνευση Θέματος ή άλλη σύνθετη διαδικασία.

Οι μεθοδολογίες που μελετήσαμε στα πλαίσια της Ανίχνευσης Τάσεων από το διαδίκτυο και τα κοινωνικά δίκτυα, εφαρμόζονται πάνω σε δεδομένα που αποτελούνται κατά κύριο λόγο από φυσική γλώσσα. Επομένως συναντάται συχνά η **Επεξεργασία Φυσικής Γλώσσας** του κειμένου (Natural Language Processing – NLP), η οποία περιλαμβάνει απλούς τρόπους γλωσσολογικής επεξεργασίας ανάλογα με τις ανάγκες της μεθόδου, και γνωστές τεχνικές όπως η **Απόδοση Ετικετών** στις λέξεις (POS tagging) και η **Αναγνώριση Επώνυμων Οντοτήτων** (Named-Entity Recognition-NER).

Γλωσσολογική επεξεργασία

Πρόκειται για την πρώτη, βασική επεξεργασία του κειμένου από απλούς αλγορίθμους και μπορεί να περιλαμβάνει:

I. Κατάτμηση κειμένου με χρήση κάποιου εργαλείου (segmentation system), η οποία αποτελεί συνήθως το πρώτο στάδιο επεξεργασίας για τα περισσότερα εργαλεία επεξεργασίας φυσικής γλώσσας

II. Αφαίρεση στοιχείων της γλώσσας τα οποία λειτουργούν ανασταλτικά στην μετέπειτα επεξεργασία ανάλογα με τις απαιτήσεις της μεθοδολογίας. Τα στοιχεία αυτά αφαιρούνται διότι δεν περιέχουν χρήσιμη πληροφορία ή/και η παρουσία τους δημιουργεί παρεμβολές. Παραδείγματα:

    i. Αφαίρεση περιττών λέξεων (stop-words/useless words). Πρόκειται για τις λέξεις οι οποίες είναι οι πιο συνηθισμένες στην κάθε γλώσσα και δεν έχουν κάποια σημασιολογική αξία. Δεν υπάρχει προκαθορισμένη λίστα, αλλά κάθε φορά το σύνολο καθορίζεται από τον αλγόριθμο επεξεργασίας φυσικής γλώσσας. Στην αγγλική γλώσσα, λέξεις που συχνά εμπεριέχονται στις αντίστοιχες λίστες είναι: “the”, “that”, “it”, “on”, “who”, “at”, “which” κ.ο.κ.

    ii. Αφαίρεση λέξεων ξένης γλώσσας

    iii. Αφαίρεση σημείων στίξης

    iv. Αφαίρεση συμβόλων

    v. Αφαίρεση ετικετών (hashtags). Ανάλογα με τις ανάγκες της μεθοδολογίας μπορεί τα hashtags να είναι επιθυμητά και να εξαχθούν για να χρησιμοποιηθούν ανάλογα είτε να θεωρηθούν περιττά και να αφαιρεθούν

    vi. Αφαίρεση html στοιχείων

    vii. Αφαίρεση εξωτερικών συνδέσμων

viii. Μετατροπή κειμένου σε μικρά γράμματα (lowercase)

ix. Αφαίρεση αναφορών σε χρήστες, εφόσον πρόκειται για αναρτήσεις σε κοινωνικά μέσα

III. Αναγωγή των λέξεων στη σημασιολογική τους ρίζα (stemming): Η διαδικασία κατά την οποία γίνεται αντιστοίχιση συγγενών λέξεων στη ρίζα τους. Δεν χρειάζεται απαραίτητα αυτή να ταυτίζεται με με τη μορφολογική ρίζα της λέξης, αρκεί τα διαφορετικά στοιχεία του λόγου ανεξαρτήτως χαρακτηριστικών (γένους, καταλήξεων, κλίσης, χρόνου κλπ) να αντιστοιχηθούν σημασιολογικά σε μια κοινή ρίζα

IV. Εξυγείανση κειμένου (text sanitization). Ως εξυγείανση κειμένου ορίζεται η διαδικασία του εντοπισμού και της αφαίρεσης ευαίσθητων δεδομένων από ένα σύνολο ή μια βάση δεδομένων

V. Εξαίρεση αναρτήσεων που δεν πληρούν κάποια κριτήρια ανάλογα με τις απαιτήσεις της μεθοδολογίας. Παραδείγματα κριτηρίων:

i. Εξαίρεση αναρτήσεων με λιγότερες από έναν προκαθορισμένο αριθμό λέξεις

ii. Εξαίρεση αναρτήσεων που είναι σε άλλη γλώσσα

iii. Εξαίρεση αναρτήσεων με πολλά σύμβολα

iv. Εξαίρεση αναρτήσεων που δεν περιέχουν ετικέτες (hashtags)

Επεξεργασία σε επίπεδο στοιχείων κειμένου

- **Απόδοση Ετικετών στις λέξεις**

Η απόδοση Ετικετών Μέρους του Λόγου (Part-Of-Speech Tagging - POS Tagging) γίνεται με τη βοήθεια ενός προγράμματος το οποίο κάνει προσπέλαση κειμένου σε κάποια συγκεκριμένη γλώσσα και αναθέτει σε κάθε λέξη το μέρος του λόγου στο οποίο ανήκει [27], [39], [40], [41]. Προσδιορίζει δηλαδή αν πρόκειται για ουσιαστικό, ρήμα, επίθετο, κ.ο.κ. Για παράδειγμα στο [27] η απόδοση ετικετών γίνεται με χρήση του Stanford POS tagger, μια υλοποίηση ανοιχτού κώδικα σε Java από το Πανεπιστήμιο του Stanford για την αγγλική γλώσσα [52].

- **Αναγνώριση Επώνυμων Οντοτήτων**

Η Αναγνώριση Επώνυμων Οντοτήτων (Named Entity Recognition - NER) πραγματοποιείται με τη χρήση αλγορίθμου ο οποίος αναγνωρίζει λέξεις και ακολουθίες λέξεων που είναι ονόματα με την ευρεία έννοια - ονόματα ανθρώπων και οργανισμών, τοπονύμια, ακόμα και ονομασίες γονιδίων, πρωτεϊνών και άλλες αντίστοιχα δεσμευμένες λέξεις της εκάστοτε γλώσσας. Συνήθως χρησιμοποιείται μαζί με την Απόδοση Ετικετών και άλλα εργαλεία επεξεργασίας φυσικής γλώσσας [39], [40], [41].

- **Εξαγωγή ετικετών (hashtags)**

Ενώ στην αρχή ήταν χαρακτηριστικό στοιχείο του Twitter, πλέον τα hashtags έχουν γίνει δημοφιλή στα περισσότερα κοινωνικά μέσα. Θεωρούνται ως ενδεικτικά του θέματος της ανάρτησης του χρήστη, αφού χρησιμοποιούνται συνήθως για να το τονίσουν και να τραβήξου την προσοχή απευθείας σε αυτό. Για το λόγο αυτό σε πολλές μεθοδολογίες Ανίχνευσης Τάσης ή

Ανίχνευσης Θέματος εντοπίζονται στα πλαισια της επεξεργασίας, ανάγονται σε λέξεις κλειδιά, γίνεται ομαδοποίηση βάσει αυτών, ή ανάγονται απευθείας σε θέματα ([33], [34], [42], [53]).

- **Εξαγωγή εικονιδίων (emoticons)**

Η εξαγωγή των εικονιδίων ή λέξεων που θεωρούνται ότι εκφράζουν συναισθήματα, γίνεται προκειμένου να ακολουθήσει στάδιο ανάλυσης συναισθήματος. Αλλιώς τα εικονίδια πιθανώς αφαιρούνται στα πλαίσια της γλωσσολογικής επεξεργασίας ως περιττά στοιχεία.

- **Τμηματοποίηση HTML κόμβων**

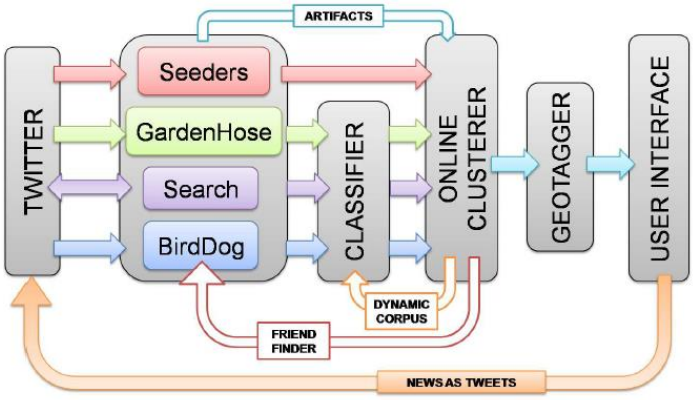
Σε μεγαλύτερα κείμενα από αυτά των κοινωνικών μέσων μικρών αναρτήσεων (microblogs), η τμηματοποίηση του κειμένου βοηθάει στην επιπλέον διαχείριση σύνθετων γλωσσικών στοιχείων όπως τα ακρονύμια, οι λέξεις που χωρίζονται με παύλες, (hyphenated words), οι αριθμοί (π.χ. ημερομηνίες, αριθμοί επικοινωνίας) κ.ο.κ.

Επεξεργασία σε επίπεδο ανάρτησης

Ανάλογα με το χαρακτήρα και τους σκοπούς της μεθοδολογίας το στάδιο της επεξεργασίας των δεδομένων μπορεί να περιλαμβάνει και επεξεργασία σε επίπεδο ανάρτησης. Η επεξεργασία μπορεί να αφορά ταξινόμηση ή ομαδοποίηση (χρονική, γεωγραφική), εξαίρεση κάποιων αναρτήσεων με βάση μη γλωσσικά κριτήρια, ή άλλη μορφή επεξεργασίας. Στη συνέχεια δίνονται παραδείγματα άλλων ειδών επεξεργασίας, μαζί με κάποια στοιχεία για τη μεθοδολογία στην οποία εντοπίζονται.

- Απόδοση γεωγραφικών ετικετών στις αναρτήσεις (GEO tagging)

Στις Εικόνες 7 και 8 φαίνεται το διάγραμμα ροής της μεθοδολογίας και η έξοδος του συστήματος Twitterstand [39] το οποίο ταξινομεί γεωγραφικά τις θεματικές ομάδες σε ένα επιπλέον στάδιο επεξεργασίας. Το σύστημα αφορά την ανίχνευση σημαντικών γεγονότων (breaking news) σε αντιστοιχία με τις σχετικές γεωγραφικές τοποθεσίες. Στο διάγραμμα ροής φαίνεται επίσης άλλο ένα ενδιάμεσο στάδιο μετά τη συλλογή των δεδομένων, το στάδιο της ταξινόμησης (classifier). Σε εκείνο το σημείο αναγνωρίζεται αν η ανάρτηση αφορά ειδησεογραφικά νέα ή όχι.



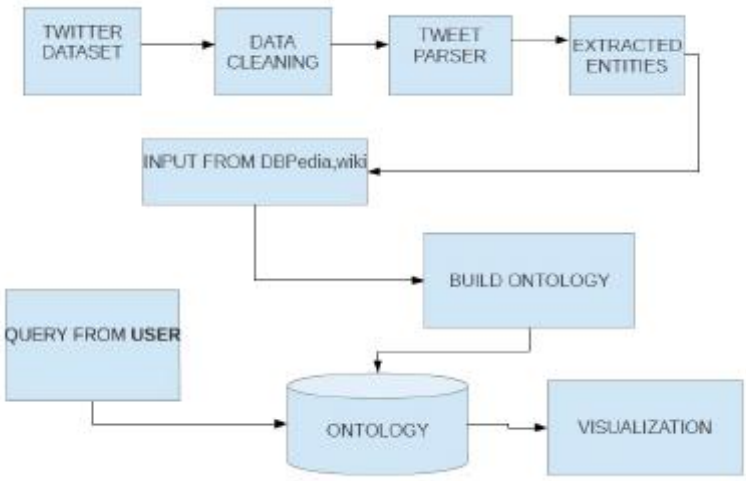
Εικόνα 7: Το διάγραμμα ροής της μεθοδολογίας πίσω από το Twitterstand [39]

Εικόνα 8: Η έξοδος του συστήματος Twitterstand και η οπτικοποίηση των αποτελεσμάτων [39]

- Χρονική ταξινόμηση ή ομαδοποίηση και καταμέτρηση αναρτήσεων

Στην Εικόνα 9 φαίνονται τα στάδια της μεθοδολογίας που χρησιμοποιεί το σύστημα Sociopedia [43], το οποίο ανιχνεύει τάσεις για προκαθορισμένα θέματα στο Twitter, με τη βοήθεια οντολογιών. Αμέσως μετά τη γλωσσολογική επεξεργασία των δεδομένων (Data Cleaning) εφαρμόζει χρονική ταξινόμηση σε αυτά και καταμέτρηση σε ημερήσια βάση.

Αυτό γίνεται στο συγκεκριμένο σύστημα με τη λογική η εφαρμογή να έχει ως έξοδο για τον τελικό χρήστη-στέλεχος μια εταιρείας την εικόνα των τάσεων σχετικά με κάποιο προϊόν σε πραγματικό χρόνο. Αυτό μπορεί να επιτευχθεί με μια γραφική παράσταση του όγκου των σχετικών αναρτήσεων σε συνάρτηση με το χρόνο.

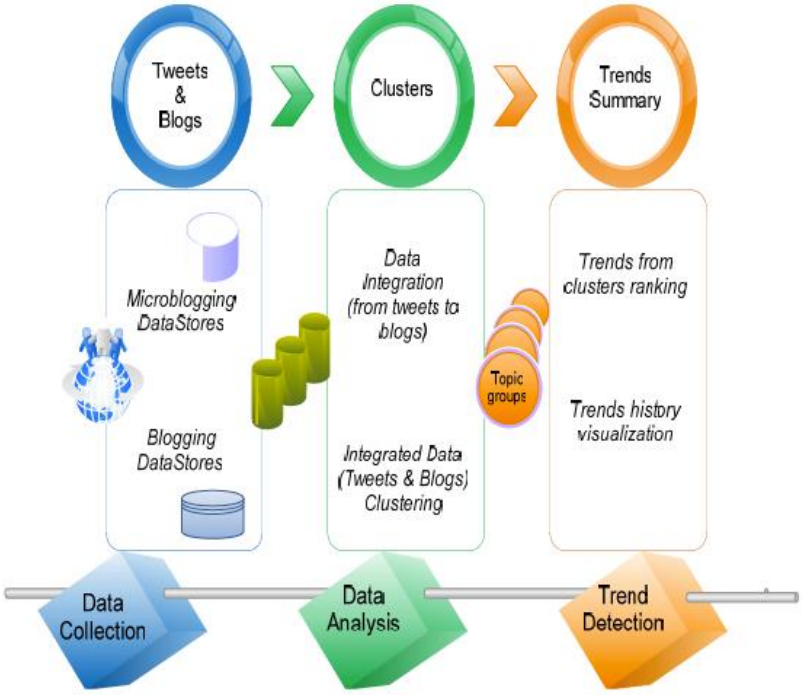


Εικόνα 9: Διάγραμμα ροής της μεθοδολογίας που χρησιμοποιεί το σύστημα Sociopedia [43]

- Παρουσίαση δεδομένων από διαφορετικές πηγές με κοινό μοντέλο (κανονικοποίηση)

Στην περίπτωση που μια μεθοδολογία ή σύστημα αφορά δεδομένα που προέρχονται από δύο ή παραπάνω διαφορετικές πηγές, είναι σκόπιμο να υπάρχει ένα στάδιο ομογενοποίησης των

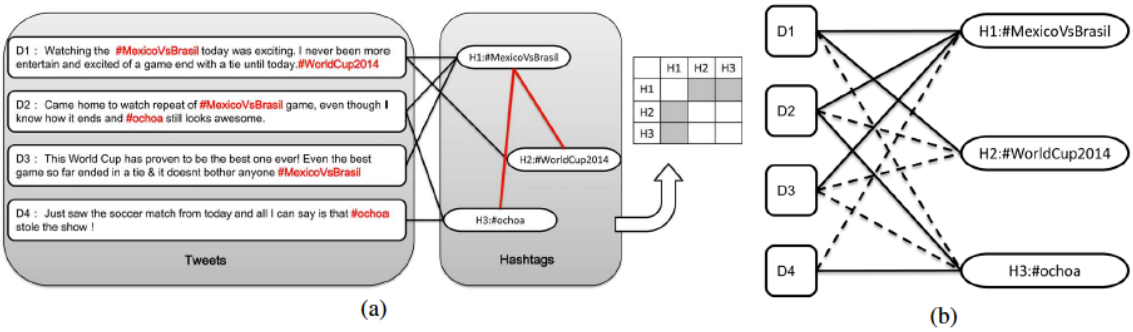
δεδομένων προκειμένου να προχωρήσουν με ενιαίο τρόπο τα επομενα στάδια της μεθοδολογίας. Για παράδειγμα το σύστημα Cloud4Trends [26] (Εικόνα 10), αντλεί δεδομένα από το Twitter και από ένα σύνολο ιστολογίων (Blogs) τα οποία επεξεργάζεται, μοντελοποιεί με κοινό τρόπο και εισάγει σε μια βάση δεδομένων προκειμένου να προχωρήσει σε ομαδοποίηση.



Εικόνα 10: Οπτική αναπαράσταση του συστήματος Cloud4trends [26]

- Εξαίρεση αναδημοσιεύσεων (retweets)

Η εξαίρεση των αναδημοσιεύσεων όταν πρόκειται για κάποιο κοινωνικό μέσο (κοινοποίηση υπάρχουσας δημοσίευσης ή “share” στο Facebook, αναδημοσίευση ή “retweet” μπορεί επίσης να αποδειχθεί χρήσιμη για τα επόμενα στάδια. Στο Hasghtag Graph based Topic Model (HGTM) [53], το συγκεκριμένο στάδιο επεξεργασίας χρησιμοποιείται μαζί με την συνήθη κανονικοποίηση δεδομένων πριν την ομαδοποίηση που κάνει το μοντέλο βάσει των hashtags (Εικόνα 11).



Εικόνα 11: Μοντέλο Ανίχνευσης Θέματος το οποίο κατά το στάδιο επεξεργασίας αποκλείει τα retweets [53]

3.3.4.3. Ανίχνευση Θέματος

Βασικές προσεγγίσεις

Προκειμένου να βγει οποιοδήποτε συμπέρασμα από ένα σύνολο κειμένων στο διαδίκτυο, είτε πρόκειται για ολόκληρα κείμενα σε κάποια ιστοσελίδα, είτε για αναρτήσεις σε ιστολόγια και κοινωνικά μέσα, είναι απαραίτητο να μπορεί να προσδιοριστεί το κύριο θέμα του κάθε κειμένου ή όλα τα θέματα που πραγματεύεται το κείμενο. Για το λόγο αυτό ανίχνευση του θέματος κειμένου με χρήση αλγορίθμων έχει αποτελέσει αντικείμενο έρευνας ήδη πριν την εμφάνιση των κοινωνικών μέσων, και άρα των αναρτήσεων μικρού μήκους οι οποίες παρουσιάζουν τις δικές τους ιδιαιτερότητες.

Όπως είναι αναμενόμενο, σε μια μεθοδολογία Ανίχνευσης Τάσεων, η Ανίχνευση Θέματος (Topic Detection) μαζί με τον Προσδιορισμό Τάσης (Trend Detection) που είναι το λογικά επόμενο βήμα, αποτελούν τα δύο σημαντικότερα στάδια. Όπως έχει προαναφερθεί μπορεί να γίνονται και ταυτόχρονα σαν ένα στάδιο, με τον Προσδιορισμό Τάσης φαινομενικά να προηγείται. Αυτό συμβαίνει σε πολλές περιπτώσεις όπου τα θέματα που ανιχνεύονται είναι απευθείας τα δημοφιλή θέματα, δηλαδή οι τάσεις (π.χ. ομαδοποίηση δημοφιλών λέξεων-κλειδιών σε θέματα οδηγεί απευθείας στο ζητούμενο). Το στάδιο ανίχνευσης θέματος λείπει από κάποια μεθοδολογία ανίχνευσης τάσης μόνο όταν τα θέματα έχουν προσδιοριστεί από πριν, ή όταν εντοπίζονται χειροκίνητα από το χρήστη του συστήματος/ερευνητή.

Στο σημείο αυτό θα αναφέρουμε περιληπτικά τις δύο βασικές προσεγγίσεις Ανίχνευσης Θέματος που χρησιμοποιούνται στα πλαίσια μιας μεθόδου Ανίχνευσης Τάσεων ([32], [30]), ενώ ακολουθεί εκτενέστερη ανάλυσή τους στα επόμενα υποκεφάλαια:

- **Πιθανοτικά Μοντέλα Θέματος**

Τα Μοντέλα Θέματος (Probabilistic Topic Models) είναι πιθανοτικοί αλγόριθμοι που στοχεύουν στον εντοπισμό της κρυμμένης θεματικής δομής σε μεγάλο αριθμό αρχείων. Τα μοντέλα LDA, pLSA και παραλλαγές τους είναι ιδιαίτερα δημοφιλή, ενώ τα τελευταία χρόνια έχουν γίνει και προσπάθειες προσαρμογής των μοντέλων θέματος στις ιδιαίτερες απαιτήσεις των αναρτήσεων μικρού μήκους και τα υπόλοιπα χαρακτηριστικά των κοινωνικών μέσων ([23], [32], [53], [34], [45]).

- **Ομαδοποίηση**

Η Ομαδοποίηση ή Ανάλυση Συστάδων (Clustering) αναφέρεται στη δημιουργία ομάδων όμοιων οντοτήτων, και χρησιμοποιείται συχνά και σε πολλές διαφορετικές μορφές προκειμένου να διαμορφωθούν οι θεματικές ενότητες ενός ή περισσότερων κειμένων. Υπάρχουν μέθοδοι ομαδοποίησης **βασισμένες σε λέξεις-κλειδιά (keyword-based approaches),** όπου η ανίχνευση θέματος πραγματοποιείται σύμφωνα με τη συχνότητα κοινής εμφάνισής τους και συνήθως προηγείται υπολογισμός της σημαντικότητας/βάρους των λέξεων-κλειδιών ή δημοφιλών όρων ([30], [31], [17], [38], [34]). Εναλλακτικά, οι μέθοδοι **βασισμένες σε χαρακτηριστικά (feature-based approaches)** επικεντρώνονται σε χαρακτηριστικά στοιχεία των αναρτήσεων (π.χ. hashtags, χρονική σήμανση, χρήστες) και κάνουν μια ομαδοποίηση των χαρακτηριστικών αυτών και όχι των αναρτήσεων απευθείας. Ομαδοποίηση μπορεί να έχουμε επίσης σε επίπεδο ανάρτησης (ομοειδείς αναρτήσεις - [42], [45], [26], [39]) ή και σε επίπεδο θέματος όπου υπο-θέματα που έχουν ανιχνευθεί ομαδοποιούνται σε ένα μεγαλύτερο ([40], [34], [41]). Τέλος, σε κοινωνικά δίκτυα που

χρησιμοποιούνται πολύ τα hashtags όπως το Twitter, μπορεί να γίνει και ομαδοποίηση βασισμένη σε αυτά.

Μοντέλα θέματος

Μια από τις συνηθέστερες προσεγγίσεις για Ανίχνευση Θέματος είναι τα Μοντέλα Θέματος (topic models). Σε αυτά, κάθε κείμενο θεωρείται ως ένα μείγμα θεμάτων, και κάθε θέμα μπορεί να αναπαρασταθεί από ένα σύνολο σημασιολογικά σχετιζόμενων λέξεων, δηλαδή μια κατανομή λέξεων από το λεξιλόγιο της κάθε γλώσσας. Στη συνέχεια χρησιμοποιούνται στατιστικές μέθοδοι για να προσδιοριστούν τα συστατικά στοιχεία του θέματος (topic-word distributions) και η σύνθεση του κειμένου (topic proportions). Στην ουσία, τα μοντέλα θέματος εντοπίζουν τα θέματα ενός κειμένου απεικονίζοντας τα μοτίβα της εμφάνισης των λέξεων σε συστάδες (co-occurrence patterns) σε επίπεδο αρχείου κειμένου [30].

Η προσπάθεια σημασιολογικής μοντελοποίησης των αρχείων κειμένου ξεκίνησε με το LSA (Latent Semantic Analysis) μοντέλο [31]. Η PLSA (Probabilistic Latent Semantic Analysis) ανάλυση [32] αποτελεί εξέλιξη του προηγούμενου, όπου το κείμενο αναπαρίσταται σαν μείγμα θεμάτων και το θέμα σαν μια πιθανοτική κατανομή λέξεων.

Ένα από τα σημαντικότερα μεταγενέστερα μοντέλα είναι η λανθάνουσα κατανομή Dirichlet (Latent Dirichlet Allocation (LDA) [33]. Το LDA είναι ένα μοντέλο στο οποίο τα λανθάνοντα θέματα βρίσκονται μετά από εκτίμηση πιθανοτικών κατανομών σε ένα σύνολο δεδομένων εκπαίδευσης (training data set). Ο LDA είναι μια τεχνική μη επιβλεπόμενης τεχνητής νοημοσύνης καθώς και ένα μοντέλο παραγωγής αρχείων προτύπων, το οποίο μπορεί να χρησιμοποιηθεί για να εντοπίσει το κρυμμένο θέμα σε μια συλλογή αρχείων.

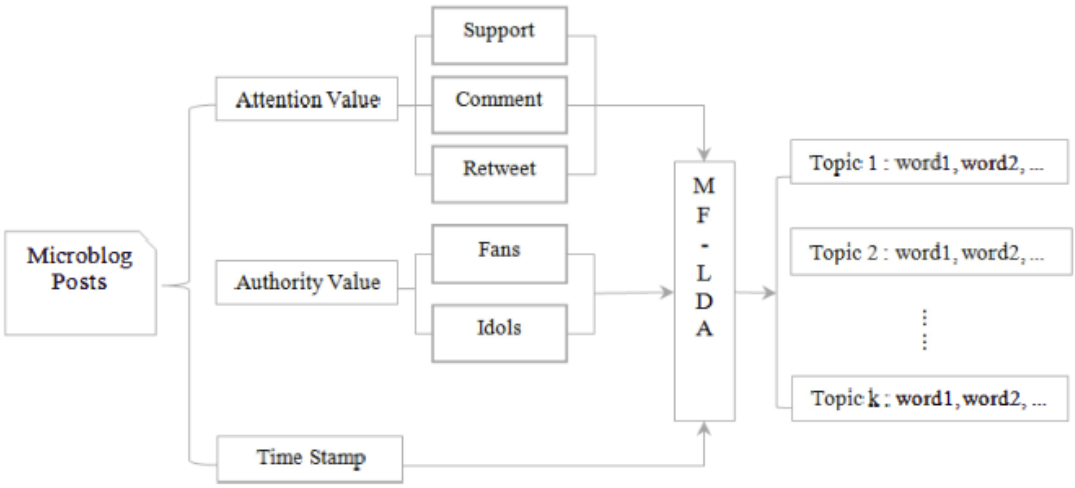
Την τελευταία δεκαετία τα Μοντέλα Θέματος έχουν μελετηθεί αρκετά καθώς έχουν χρησιμότητα για τις περισσότερες εφαρμογές της Εξόρυξης Δεδομένων, και συγκεκριμένα για το LDA έχουν προταθεί αρκετές ενδιαφέρουσες παραλλαγές [34].

Στην περίπτωση της εφαρμογής των δημοφιλών μοντέλων στα κοινωνικά μέσα, προκύπτει ένα βασικό πρόβλημα λόγω της συντομίας των αναρτήσεων (short texts/microblogging posts). Το πρόβλημα αυτό αναφέρεται στη βιβλιογραφία ως πρόβλημα αραιών δεδομένων (data sparsity problem), ή πρόβλημα αραιών συνδέσμων (sparse context links problem), και συνοψίζεται στο ότι οι συγγενείς όροι σπάνια θα αναφέρονται μαζί σε μια ανάρτηση, λόγω του μικρού μεγέθους της. Αν θεωρήσουμε δηλαδή κάθε ανάρτηση σαν ένα αρχείο, δεν θα μπορούμε να εντοπίσουμε αποτελεσματικά μοτίβα κοινής εμφάνισης όρων και άρα να υποδείξουμε τη λανθάνουσα θεματική δομή.

Για την αντιμετώπιση αυτού του φαινομένου, έχουν προταθεί προσαρμοσμένες εκδοχές των μοντέλων θέματος, προσεγγίσεις που βασίζονται στην ενοποίηση πολλών αναρτήσεων μικρού μήκους σε ψευδο-αρχεία για την εκπαίδευση των μοντέλων, αλλά και προσπάθειες εκμετάλλευσης εξωτερικών πηγών (external knowledge) όπως μεγαλύτερα κείμενα σχετικού περιεχομένου.[30]

Στη συνέχεια αναφέρουμε ενδεικτικά Μοντέλα Θέματος που εντοπίστηκαν κατά την βιβλιογραφική επισκόπηση σε μεθοδολογίες Ανίχνευσης Τάσης στο στάδιο Ανίχνευσης Θέματος:

- **LDA (Latent Dirichlet Allocation):** Χρησιμοποιείται όπως όλα τα κλασικά μοντέλα θέματος για τον εντοπισμό του θέματος σε οποιοδήποτε αρχείο κειμένου (ιστολόγια, άρθρα, αναρτήσεις σε κοινωνικά δίκτυα).
- **MF-LDA (Microblog Features LDA):** Παραλλαγή του κλασικού LDA μοντέλου θέματος [32] η οποία ενσωματώνει τεχνική ομαδοποίησης βασισμένη σε 5 χαρακτηριστικά της ανάρτησης. Η παραλλαγή διευκολύνει τον εντοπισμό θέματος σε μικρά κείμενα (κοινωνικά δίκτυα-microblogs).
- **HGTM (Hashtag Graph Topic Model):** Παρόμοιο μοντέλο με το LDA στη λειτουργία του, που χρησιμοποιεί ομαδοποίηση βάσει των ετικετών (hashtags) για καλύτερη προσαρμογή σε μικρά κείμενα (κοινωνικά δίκτυα-microblogs) και στο σύγχρονο τρόπο σύνταξης των αναρτήσεων.
- **PLDA (Partially Labeled Dirichlet Allocation**): Παραλλαγή του LDA μοντέλου με μερική προσήμανση των αναρτήσεων.
- **PAM (Pachinko Allocation Model):** Εναλλακτικό μοντέλο θέματος που χρησιμοποιεί ακυκλικό κατευθυνόμενο γράφο (Directed Acyclic Graph-DAG) για να εντοπίσει συσχετισμούς μεταξύ θεμάτων, επεκτείνοντας την έννοια του θέματος σαν κατανομή λέξεων αλλά και υπο-θεμάτων (subtopics).

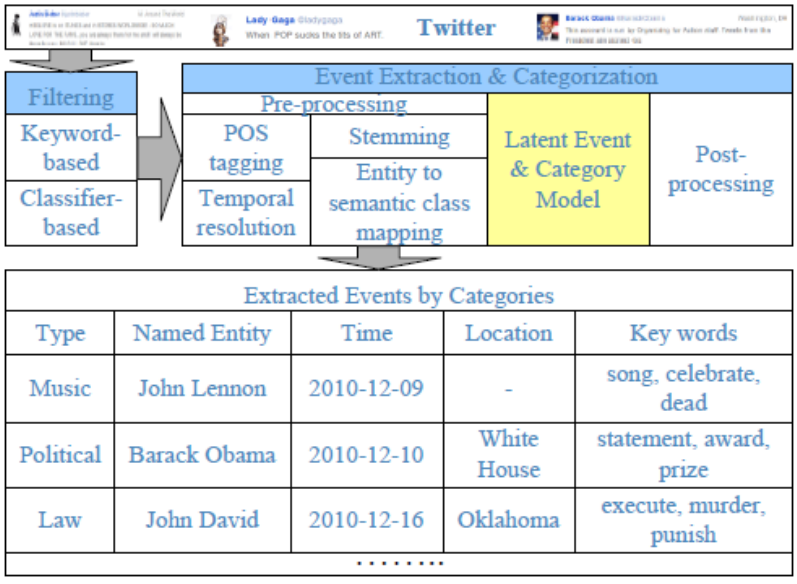


Εικόνα 12: Γραφική απεικόνιση του MF-LDA Μοντέλου Θέματος σαν στάδιο Ανίχνευσης θέματος/τάσης [32]

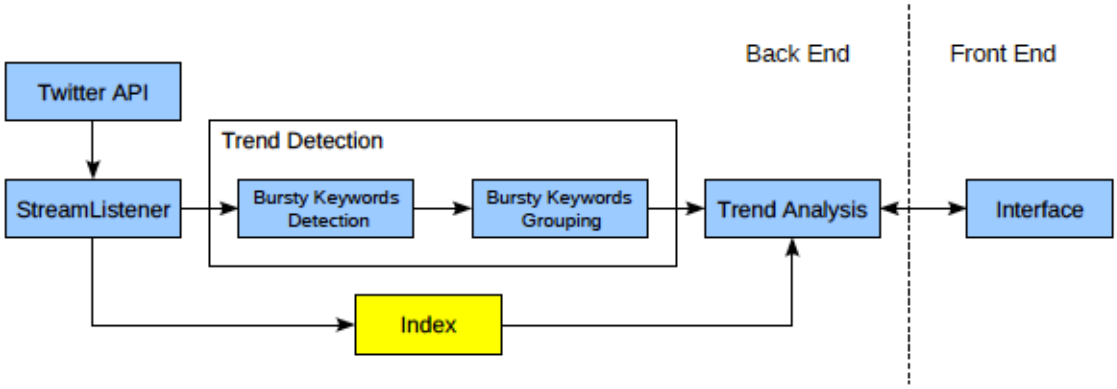
Ομαδοποίηση

Σαν δεύτερη κύρια εναλλακτική τεχνική στο στάδιο της Ανίχνευσης Θέματος, η Ομαδοποίηση ή Ανάλυση Συστάδων στοχεύει στο να συγκεντρώσει τους συγγενείς νοηματικά όρους, διαμορφώνοντας έτσι οικογένειες όρων, που τελικά είναι τα θέματα των κειμένων ή των αναρτήσεων. Στα κοινωνικά μέσα είναι ιδιαίτερα δημοφιλής τεχνική καθώς δεν περιορίζεται από το πρόβλημα των αραιών συνδέσμων (sparse context links problem). Κάποια είδη ομαδοποίησης ή ανάλυσης συστάδων που εντοπίστηκαν να χρησιμοποιούνται κατά την έρευνα είναι:

- Ομαδοποίηση λέξεων κλειδιών (keyword clustering) [31], [35], [34], [30]
- Ομαδοποίηση βασισμένη σε ταυτόχρονη εμφάνιση (clustering based on co-occurences) [31], [17], [38]
- Ομαδοποίηση βάσει κατανομής (clustering by distribution) [17], [38]
- Ομαδοποίηση βασισμένη σε χαρακτηριστικά (feature/attribute based clustering) [26], [32]
- Ομαδοποίηση Gaussian (online clustering Gaussian) [26]
- Ομαδοποίηση σε επίπεδο ανάρτησης (document-based clustering) [39], [26]
- Ομαδοποίηση θεμάτων [40], [34], [41]
- Ομαδοποίηση βάσει γεωγραφικής τοποθεσίας (geographical clustering) [39], [35]
- Ομαδοποίηση σε επίπεδο φράσης-κλειδιού [36]
- Ομαδοποίηση βάσει ομοιότητας [36]
- Ομαδοποίηση βάσει της σημασιολογικής απόστασης [42]
- Ομαδοποίηση βάσει των hashtags [53], [33]



Εικόνα 13: Ομαδοποίηση σε επίπεδο θέματος με το LECM μοντέλο [41]

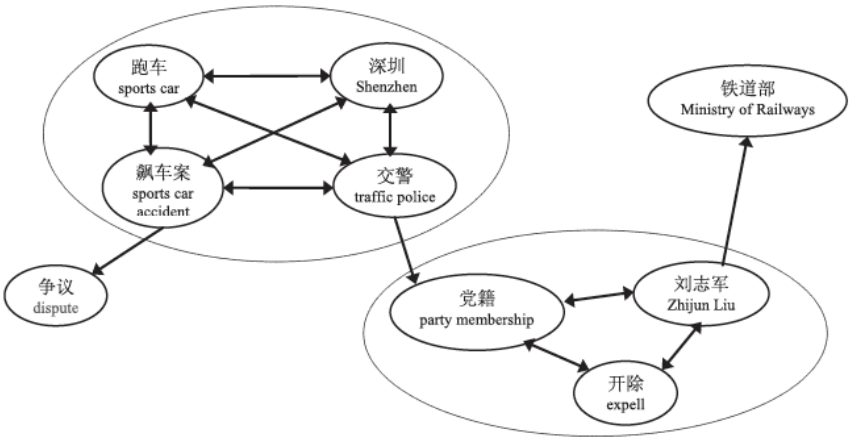


Εικόνα 14: Ομαδοποίηση λέξεων-κλειδιών βάσει κοινής εμφάνισης (Twittermonitor) [31]

Μέθοδοι ανάλυσης γράφων

Ο εντοπισμός των θεμάτων μπορεί επίσης να γίνει με τη βοήθεια γράφων και μεθόδων ανάλυσης γράφων (Graph-based approaches). Στην σχετικά πρόσφατη βιβλιογραφία, υπάρχει ποικιλία αλγορίθμων που τοποθετούν τα επιθυμητά στοιχεία (όρους, αναρτήσεις) σε διάταξη γράφου και εκμεταλλεύονται τις ιδιότητες της διάταξης για τον εντοπισμό των μοτίβων που οδηγούν στην ανίχνευση θέματος. Χαρακτηριστικό παράδειγμα της εν λόγω χρήσης είναι η μετατροπή των λέξεων-κλειδιών σε γράφο με κατάλληλο αλγόριθμο με βάση την κοινή τους εμφάνιση, και κατάτμηση του γράφου σε υπογράφους οι οποίοι αντιστοιχούν σε θεματικές ενότητες (Αλγόριθμοι KeyGraph [60], IdeaGraph [61]).

Στο [38], δημιουργείται ένα γράφος με κόμβους τις ανερχόμενες λέξεις κλειδιά που έχουν εντοπιστεί, και οι ακμές του γράφου βαθμονομούνται επίσης ανάλογα με το βαθμό συσχέτισης των κόμβων που συνδέουν. Εφόσον το θέμα ορίζεται σαν ένα σύνολο σημασιολογικά σχετιζόμενων λέξεων, σταδιακά αφαιρούνται οι ακμές των λέξεων που βρίσκονται πολύ κοντά. Για την ανεύρευση των ανερχόμενων θεμάτων σε χρονικό t, γίνεται αναζήτηση των υπογράφων που προκύπτουν κατ'αυτό τον τρόπο. Όλες οι αναρτήσεις που περιλαμβάνουν τους όρους του υπογράφου, θεωρείται ότι ανήκουν στη θεματική του ενότητα.



Εικόνα 15: Σχηματική αναπαράσταση των υπογράφων που σχηματίζουν οι ανερχόμενες λέξεις-κλειδιά [38]

Άλλοι αλγόριθμοι

Άλλοι αλγόριθμοι που έχουμε δει να χρησιμοποιούνται συμπληρωματικά στην Ανίχνευση Θέματος:

1. Αλγόριθμος K-Means
2. Αλγόριθμος K-NN
3. Αλγόριθμος Viterbi

3.3.4.4. Προσδιορισμός τάσης

Όπως έχουμε δει, οι μεθοδολογίες που έχουν αναπτυχθεί με σκοπό την ανίχνευση τάσεων, μπορεί να αναφέρονται σε εντοπισμό Ανερχόμενων Θεμάτων, εντοπισμό Σημαντικών Γεγονότων ή Ανίχνευση Τάσεων για προκαθορισμένα θέματα. Ανάλογα με το τι είδους «τάση» αναζητείται, διαφοροποιούνται αρκετά τα στάδια Ανίχνευσης Θέματος και Προσδιορισμού Τάσης. Με γνώμονα αυτά αλλά και την απευθείας ανίχνευση δημοφιλών θεμάτων (συνύπαρξη των δύο σταδίων), μπορούμε να ξεχωρίσουμε τρεις κατηγορίες:

Προσδιορισμός τάσης χωρίς να έχει προηγηθεί ανίχνευση θέματος

Αναφερόμαστε στις περιπτώσεις όπου στη μεθοδολογία δε συμπεριλαμβάνεται στάδιο ανίχνευσης θέματος. Τα θέματα είναι προκαθορισμένα από πριν και διευερνάται η πορεία τους ή το κατά πόσο είναι δημοφιλή σε μια δεδομένη χρονική στιγμή. Τεχνικές:

- Στοχαστικά μοντέλα (π.χ. Hawkes process)
- Χρήση οντολογιών και μέτρηση συχνότητας εμφάνισης
- Αλγόριθμοι πλησιέστερου γείτονα (π.χ. K-NN)

Προσδιορισμός τάσης ενώ έχει προηγηθεί ανίχνευση θέματος

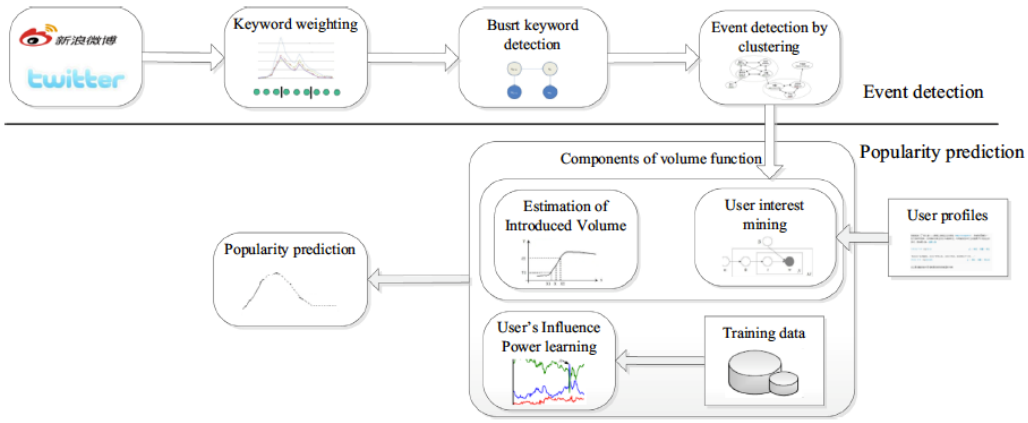
Στις περιπτώσεις που τα δύο στάδια είναι ξεχωριστά, έχει προηγηθεί ο εντοπισμός των θεμάτων και το επόμενο στάδιο αφορά το αν το κάθε θέμα είναι ανερχόμενο (τάση) ή όχι. Τότε στο στάδιο εντοπισμού τάσης μπορεί να έχουμε:

- Χειροκίνητο εντοπισμό: Ο χρήστης αποφασίζει αν τα θέματα που έχουν εντοπιστεί είναι σημαντικά ή όχι
- TF-IDF προσέγγιση προσαρμοσμένη σε επίπεδο θέματος
- Στοχαστικό μοντέλο που προβλέπει την μελλοντική εξέλιξη (τάση) του κάθε θέματος
- Στατιστική ανάλυση/υπολογισμό παραμέτρων του θέματος και έλεγχο κατωφλίων που έχουν προσδιοριστεί
- Στατιστική παρακολούθηση του θέματος στο χρόνο (σύγκριση με προηγούμενες περιόδους)

Στις περιπτώσεις που το στάδιο ανίχνευσης θέματος συνυπάρχει με τον προσδιορισμό τάσης, προηγείται η ανίχνευση τάσης μέσω του εντοπισμού δημοφιλών ή ανερχόμενων όρων και στη συνέχεια οι όροι αυτοί (λέξεις, λέξεις-κλειδιά, φράσεις, hashtags) ομαδοποιούνται σε θέματα. Τρόποι με τους οποίους εντοπίζονται οι δημοφιλείς όροι:

- **Απόδοση βάρους στις λέξεις (word weighting)**

Η συγκεκριμένη τεχνική αποτελεί έναν τρόπο ουσιαστικής προετοιμασίας των δεδομένων για το στάδιο της Ανίχνευσης Τάσης. Χρησιμοποιείται σε μεθοδολογίες οι οποίες ανιχνεύουν δημοφιλείς όρους, και συγκεκριμένα δημοφιλείς/ανερχόμενες λέξεις στο κείμενο, προκειμένου αργότερα να τις ομαδοποιήσουν σε δημοφιλή/ανερχόμενα θέματα (π.χ. [38]). Το βάρος της κάθε λέξης υπολογίζεται με τη βοήθεια ενός ή περισσότερων αλγορίθμων για ένα ή περισσότερα χαρακτηριστικά, και ανάγεται σε ενιαία κλίμακα (π.χ. από το 0 έως το 1). Για παράδειγμα, το βάρος μιας λέξεις που εμφανίζεται περισσότερο στο κείμενο της ανάρτησης (TF, term frequency) και εμφανίζεται σε χρήστες με μεγαλύτερη επιρροή με βάση κάποιον PageRank αλγόριθμο θα είναι μεγαλύτερο από αυτό μιας λέξης που δεν εμφανίζεται συχνά ή εμφανίζεται σε αναρτήσεις χρηστών με λίγους ακολούθους. Οι λέξεις με μεγάλο βάρος, αποτελούν τις ανερχόμενες λέξεις.



Εικόνα 16: Σύστημα που χρησιμοποιεί απόδοση βάρους στις λέξεις [38]

- **TF-IDF (Term Frequency – Inverse Document Frequency)**

Στατιστική μέθοδος που χρησιμοποιεί την απόδοση βάρους και προσδιορίζει πόσο σημαντική είναι μια λέξη μέσα σε ένα αρχείο ή ένα σύνολο αρχείων. Εντοπίζονται οι ανερχόμενοι όροι και ακολουθεί ομαδοποίηση. Η μέθοδος αποδίδει βάρος στον κάθε όρο με βάση δύο μέτρα: (1) τη συχνότητα εμφάνισης του όρου μέσα στο κείμενο, και (2) τον αριθμό των κειμένων τα οποία περιέχουν τον όρο [17]. Ο γενικός τύπος της TF-IDF προσέγγισης είναι:

$$w(t_i) = tf(t_i, d_j) * \log 2 \frac{N}{df(t_i)},$$

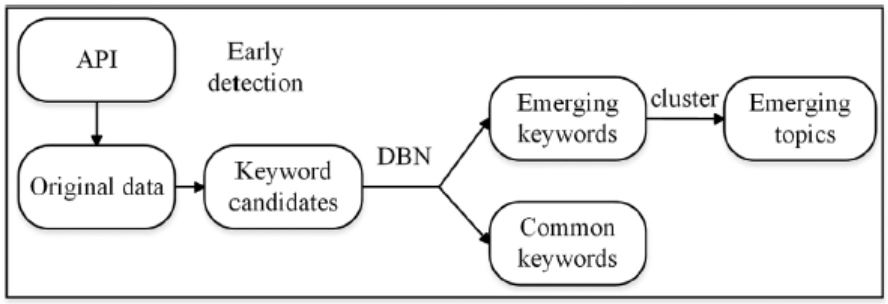
όπου N είναι ο τελικός αριθμός των κειμένων/αναρτήσεων, tf(ti, dj) η συχνότητα εμφάνισης του όρου ti μέσα στο κείμενο dj, και df(ti) είναι ο αριθμός των κειμένων μέσα στο σύνολο που περιλαμβάνει τον όρο ti [59].

Στην περίπτωση των αναρτήσεων μικρού μήκους των μέσων κοινωνικής δικτύωσης, η συχνότητα εμφάνισης του όρου στο κείμενο μειώνεται δραστικά. Πρόκειται για πρόβλημα παρόμοιο με αυτό των αραιών συνδέσμων περιεχομένου που είδαμε ότι παρουσιάζεται στα μοντέλα θέματος. Από την άλλη, αν γίνει συνένωση όλων των αναρτήσεων σε ένα ενιαίο αρχείο, γίνεται μονάδα ο παρονομαστής του τύπου.

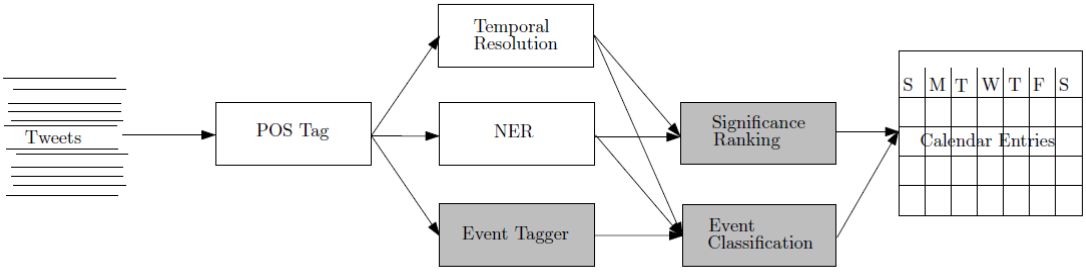
Για να ξεπεραστεί το πρόβλημα, όταν τα δεδομένα προέρχονται από μέσα με αναρτήσεις μικρού μήκους χρησιμοποιείται κάποια παραλλαγή της βασικής μεθόδου, η οποία χρησιμοποιεί υποσύνολα των αναρτήσεων τα οποία αντιμετωπίζει σαν ενιαία μεγαλύτερα κείμενα. Ο διαχωρισμός σε υποσύνολα μπορεί να γίνει είτε με βάση τη χρονική εγγύτητα συλλογής των αναρτήσεων [17], είτε με βάση τον κοινό όρο που περιέχουν [27], [42] ή την ομοιότητά τους με βάση κάποιο άλλο χαρακτηριστικό.

- **DBN (Dynamic Bayesian Network)**

Αλγόριθμος για τον εντοπισμό ανερχόμενων λέξεων-κλειδιών (trending keywords). Στη συνέχεια οι ήδη χαρακτηρισμένες ως ανερχόμενες λέξεις ομαδοποιούνται (clustering) σε θέματα με τη βοήθεια του αλγορίθμου DBSCAN.



Εικόνα 17: Χρήση DBN μοντέλου σε γράφο για τον εντοπισμό ανερχόμενων λέξεων [30]



3.3.4.5. Ανάλυση συναισθήματος

Η **Ανάλυση Συναισθήματος** (Sentiment Analysis - SA), η οποία συχνά αποκαλείται και **Εξόρυξη Γνώμης** (Opinion Mining) είναι μια ποσοτική μελέτη των απόψεων, συναισθημάτων και της αντιμετώπισης που εκφράζεται σε ένα κείμενο και σχετίζεται με κάποια οντότητα [62]. Εναλλακτικά, πρόκειται για τη διαδικασία της ανίχνευσης, εξόρυξης και ταξινόμησης των απόψεων για ένα ή περισσότερα θέματα, όπως εκφράζεται στο κείμενο ([63], [64]). Τεχνικές ανάλυσης συναισθήματος χρησιμοποιούνται σε περιπτώσεις όπου απαιτείται η αναγνώριση της κοινής γνώμης σε σχέση με κάποιο πολιτικό ζήτημα, για ζητήματα επιχειρηματικής ευφυΐας, για τη μέτρηση της ικανοποίησης του καταναλωτικού κοινού, την πρόβλεψη πωλήσεων για κάποιο προϊόν, και σε πολλές ακόμα εφαρμογές.

Με την άνοδο του ηλεκτρονικού εμπορίου (e-commerce), τα συναισθήματα, οι αξιολογήσεις και οι κριτικές του καταναλωτικού κοινού εκφράζονται ολοένα και περισσότερο δημόσια στο διαδίκτυο. Οι περισσότεροι χρήστες προκειμένου να αγοράσουν ένα προϊόν ή να εμπιστευτούν κάποιο κατάστημα, λαμβάνουν υπόψιν τις γνώμες των άλλων χρηστών που είναι διαθέσιμες σε αυτούς. Το ίδιο ισχύει και για τις υπηρεσίες, καθώς οι απόψεις που εκφράζονται σε ιστοσελίδες όπως Amazon, IMDb, epinions.com, κ.α. μπορούν να επηρεάσουν τις αποφάσεις για αγορά υπηρεσιών ή συνδρομή σε αυτές [65].

Ταυτόχρονα, η ευρεία χρήση των κοινωνικών μέσων ενθαρρύνει τους περισσότερους χρήστες να εκφράζουν μέσα από αυτά τις απόψεις τους για όλα τα θέματα που τους απασχολούν, συμμετέχοντας διαδραστικά σε συζητήσεις πολιτικού, κοινωνικού ή προσωπικού ενδιαφέροντος [63]. Επομένως τα μέσα κοινωνικής δικτύωσης αποτελούν άλλο ένα πεδίο εφαρμογής της ανάλυσης συναισθήματος που μπορεί να χρησιμεύσει στη μελέτη της συμπεριφοράς του καταναλωτή και τον προσδιορισμό των «συναισθηματικών» μοτίβων που κυριαρχούν στην αγορά. Για το λόγο αυτό η Ανάλυση Συναισθήματος μπορεί να εξυπηρετήσει και την Ανίχνευση Τάσεων της αγοράς και να αποτελέσει στάδιο κάποιων μεθοδολογιών, αν και δεν αποτελεί προαπαιτούμενο για ένα ολοκληρωμένο σύστημα Ανίχνευσης Τάσης. Στο συμπέρασμα αυτό μπορεί κανείς να καταλήξει εύκολα, παρατηρώντας ότι πολλές μεθοδολογίες ανίχνευσης τάσης δεν περιέχουν καθόλου ανάλυση συναισθήματος (π.χ. [36], [39], [31], [17], [26], [40]).

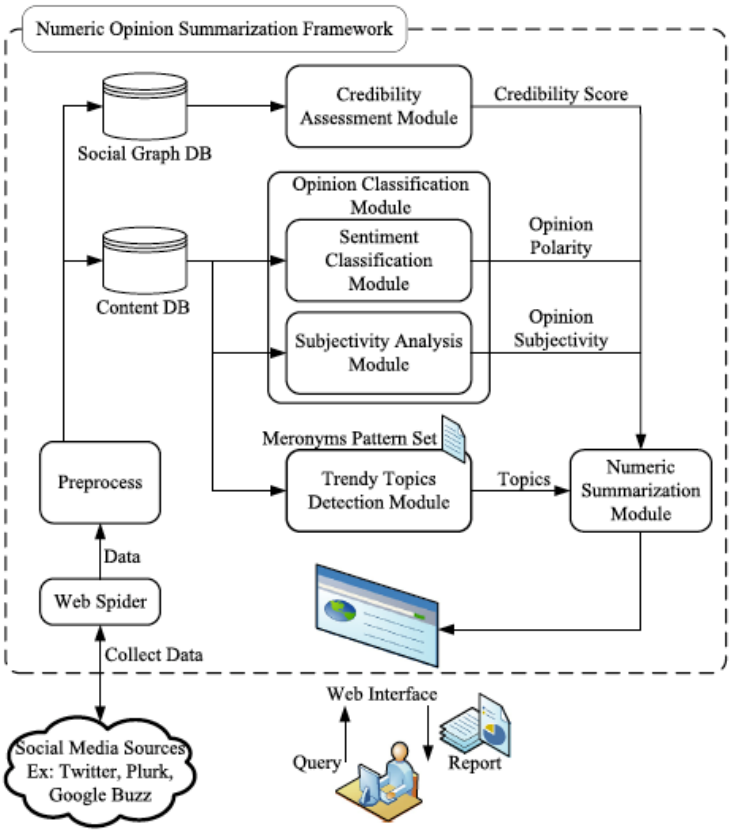
Στην περίπτωση που εμπεριέχεται στάδιο Ανάλυσης Συναισθήματος σε μια μεθοδολογία Ανίχνευσης Τάσεων, το στάδιο αυτό χρειάζεται για να προσδιοριστεί αν η τάση που εντοπίζεται έχει κάποιο συναισθηματικό προσανατολισμό [23].

Η ανάλυση συναισθημάτων γίνεται με τη βοήθεια τεχχνικών επεξεργασίας φυσικής γλώσσας (NLP) αξιοποιώντας βασικές μεθόδους κατηγοριοποίησης κειμένου (οντότητες) και υπολογιστικής γλωσσολογίας για την αναγνώριση και εξαγωγή υποκειμενικής πληροφορίας από διάφορες πηγές. Σκοπός είναι να προσδιοριστεί ο προσανατολισμός (θετικός, αρνητικός, ουδέτερος) της γνώμης που εκφέρεται για την εκάστοτε οντότητα τη συγκεκριμένη χρονική στιγμή. Αυτό γίνεται με δύο κυρίως προσεγγίσεις:

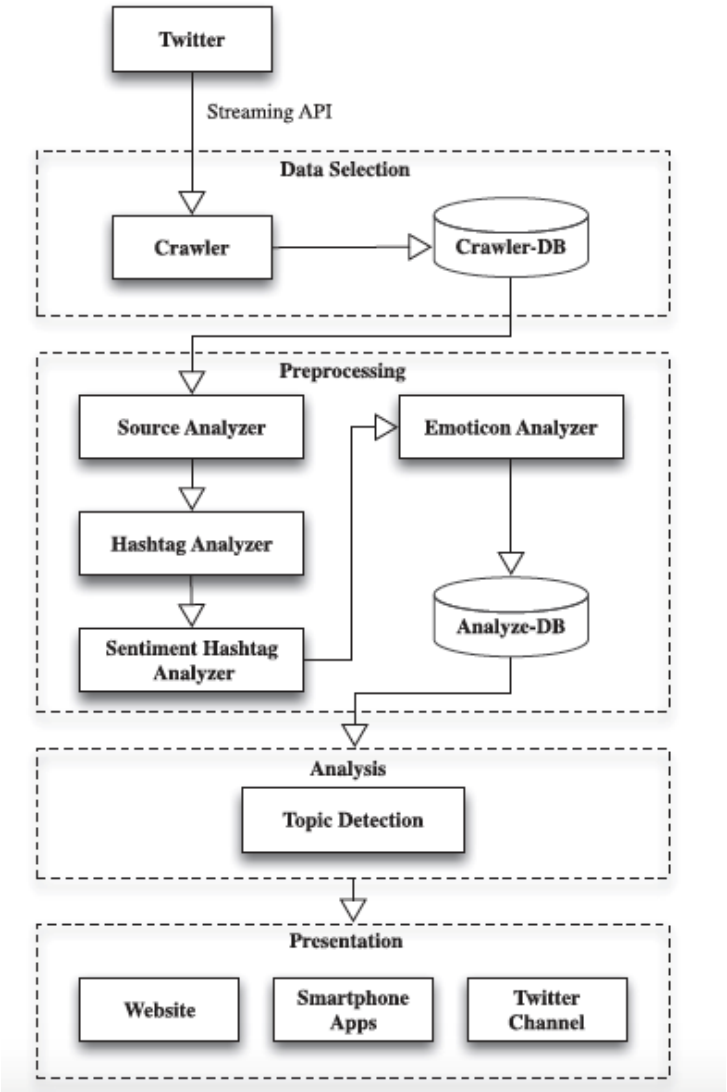
1. Ανάπτυξη λεξικών με βαθμολογημένες λέξεις ως προς τον συναισθηματικό τους προσανατολισμό, και στη συνέχεια κατηγοριοποίηση του κειμένου ανάλογα με τις πηγές αυτές. Ένα τέτοιο «συναισθηματικό λεξικό» ελεύθερο προς χρήση είναι το SentiWordNet [21].

2. Χρήση μεθόδων μηχανικής μάθησης για τη δημιουργία ενός μοντέλου που θα κατηγοριοποιεί τα συναισθήματα στο κείμενο. Υπάρχουν διάφορες μέθοδοι, όπως το SVM

μοντέλο (support vector machine), και το naïve Bayes [18]. Αυτές οι μέθοδοι ανήκουν στην ευρύτερη κατηγορία των επιβλεπόμενων μεθόδων μηχανικής μάθησης, ενώ η ομαδοποίηση/ανάλυση συστάδων που έχουμε δει σε προηγούμενα κεφάλαια αποτελεί μη-επιβλεπόμενη μέθοδο.



Εικόνα 18: Ολοκληρωμένο σύστημα Ανίχνευσης Τάσεων στα κοινωνικά μέσα που περιλαμβάνει στάδιο ανάλυσης συναισθήματος [27]



Εικόνα 19: Σύστημα Ανίχνευσης Τάσεων στο Twitter που περιλαμβάνει στάδιο ανάλυσης συναισθήματος [33]

4. Ανίχνευση τάσεων της αγοράς στην βιομηχανία κατασκευής παιχνιδιών

4.1. Παρουσίαση κλάδου

4.1.1. Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται η αγορά των παραδοσιακών και συμβατικών παιχνιδιών καθώς και η θέση της βιομηχανίας παιχνιδιών στην Ερωπαϊκή Ένωση (ΕΕ) και στις τρίτες χώρες.

Η ΕΕ διαθέτει τη μεγαλύτερη ενιαία αγορά για αγαθά και υπηρεσίες παγκοσμίως. Εκτιμάται ότι η αγορά παραδοσιακών παιχνιδιών και παιχνιδιών της ΕΕ, συμπεριλαμβανομένης της Κροατίας, ανήλθε σε 15,8 δισ. ευρώ το 2011 σε τιμές λιανικής πώλησης. Σε σύγκριση, η αμερικανική αγορά ακολουθεί με 14 δις. ευρώ. Η κινεζική αγορά παρουσίασε πωλήσεις ύψους 4,8 δισ. ευρώ, με προδιάθεση για υψηλή δυναμική εάν τα επίπεδα του εισοδήματος της χώρας συνεχίσουν να αυξάνονται *(Πηγή: European Consortium, "Study on the competitiveness of the toy industry," Rotterdam, 2013).*

Παρόλο που πρόκειται για έναν σταθερά κερδοφόρο κλάδο, αρκετοί εξωτερικοί παράγοντες ενδέχεται να επηρεάσουν αρνητικά τη ζήτηση παραδοσιακών παιχνιδιών τα επόμενα χρόνια. Το ένα είναι η γήρανση της κοινωνίας στις ώριμες αγορές, που παρατηρείται ήδη από τον περίπου σταθερό αριθμό παιδιών στην ΕΕ και τις ΗΠΑ. Στην Κίνα, η πολιτική του ενός παιδιού και η αύξηση των εισοδημάτων έχουν οδηγήσει σε απότομη πτώση του παιδικού πληθυσμού. Ένας άλλος παράγοντας είναι ο αυξημένος ανταγωνισμός των νέων προϊόντων τεχνολογίας που γίνονται στενά υποκατάστατα των παραδοσιακών παιχνιδιών: τα βιντεοπαιχνίδια, τα έξυπνα τηλέφωνα, τα "tablets" και άλλα προϊόντα ψυχαγωγίας ανταγωνίζονται τα παραδοσιακά παιχνίδια στις προτιμήσεις των παιδιών και την επακόλουθη δαπάνη των γονέων στις ώριμες και αναδυόμενες αγορές. Τα ηλεκτρονικά παιχνίδια, όπως οι εφαρμογές για "tablets", είναι άμεσα και φθηνά υποκατάστατα των προσχολικών παιχνιδιών.

Σε μακροοικονομικό επίπεδο, η αβεβαιότητα σχετικά με την ανάκαμψη της ΕΕ και των ΗΠΑ από την παγκόσμια κρίση και οι συνέπειες για τις συνθήκες ζήτησης στις αναδυόμενες οικονομίες έδειχναν ότι οι προβλέψεις για την ανάπτυξη ήταν εξαιρετικά αβέβαιες. Παρόλα αυτά, η παγκόσμια κατανάλωση παιχνιδιών αναμενόταν να αυξηθεί κατά περίπου 7,5% ετησίως έως το 2016.

Το μέγεθος της βιομηχανίας παιχνιδιών της ΕΕ υπολογίζεται περίπου σε 5,8 δισεκατομμύρια ευρώ, σε τιμές κόστους παραγωγής. Η δε άμεση απασχόληση στην παραγωγή παραδοσιακών παιχνιδιών εκτμάται περίπου στις 51.000 θέσεις για την ΕΕ. Η έμμεση απασχόληση, εξαιρουμένης της λιανικής, υπολογίζεται περίπου στα ίδια επίπεδα. Η ΕΕ υπερβαίνει στην παραγωγή και την απασχόληση των παιχνιδιών στις ΗΠΑ, η οποία εκτιμάται σε 4,4 δισ. ευρώ και σε 35.000 εργαζόμενους αντίστοιχα. Η μεγαλύτερη παραγωγή παιχνιδιών πραγματοποιείται στην Κίνα, με αξία παραγωγής 16 δισ. ευρώ. Η εκτίμηση της απασχόλησης για την Κίνα, αν και αβέβαιη, εκτιμάται σε περίπου 128.000 εργαζόμενους που εμπλέκονται στην παραγωγή παραδοσιακών παιχνιδιών.

Η παραδοσιακή και συμβατική αγορά παιχνιδιών δείχνει μέτρια ποσοστά ανάπτυξης στην Ευρώπη και τις ΗΠΑ και ισχυρούς ρυθμούς ανάπτυξης στην Κίνα και ιδιαίτερα στον υπόλοιπο κόσμο. Τα επίπεδα ανάπτυξης των παραδοσιακών και των συμβατικών παιχνιδιών είναι υψηλότερα σε σχέση με την οικονομία στο σύνολό της, προσφέροντας θετικές προοπτικές με ευκαιρίες επέκτασης, ιδίως

για τους ευρωπαίους παραγωγούς παιχνιδιών, οι οποίοι είναι ο δεύτερος σημαντικότερος εξαγωγέας παιχνιδιών μετά την Κίνα. Το 2011, οι εξαγωγές της ΕΕ στο σύνολό τους ανήλθαν σε 5,3 δισ. ευρώ, εκ των οποίων το ενδοκοινοτικό εμπόριο ανήλθε σε 4,2 δισ. ευρώ.

Στον πίνακα που ακολουθεί παρουσιάζονται οι βασικές πτυχές της αγοράς και της βιομηχανίας του κλάδου.

Πίνακας 2: Τα οικονομικά μεγέθη της αγοράς παιχνιδιών (Πηγές: Eurostat, Euromonitor, Ecofys)

Χώρα	Κατανάλωση σε εκ. €	Παραγωγή σε εκ. €	Άμεση απασχόληση (εργαζόμενοι)
Ε.Ε. 28 Σύνολο	15,828.40	5,833.61	50,902
Άλλες χώρες			
Η.Π.Α.	13.971.70	4,382.33	35,037
Κίνα	4,802.80	16,011.30	128,012
Ιαπωνία	5,201.10	2,200.08	17,590

Η παραγωγή παραδοσιακών παιχνιδιών στην ΕΕ είναι ανταγωνιστική από άποψη κόστους όταν το κόστος μεταφοράς από την Κίνα είναι υψηλό. Η παραγωγή στην ΕΕ περιλαμβάνει μικρά πλαστικά αντικείμενα χαμηλού κόστους, όπου ο όγκος των παραγγελιών είναι συχνά μικρότερος από το ποσό που απαιτείται για να αντισταθμιστεί αποτελεσματικά το κόστος μεταφοράς από την Ασία. Τα παιχνίδια που παράγονται στην ΕΕ σε αυτοματοποιημένα εργοστάσια μπορούν επίσης να είναι ανταγωνιστικά ως προς τις τιμές, ιδίως εάν οι σχετικές εισροές μπορούν να αντληθούν τοπικά.

Ορισμένες επιχειρήσεις επιλέγουν σκόπιμα να παράγουν στην ΕΕ για τους λόγους αυτούς, με παραδείγματα μεγάλα ονόματα του χώρου, όπως η LEGO και η Playmobil.

Οι καταναλωτές είναι αρκετά ευαίσθητοι στις τιμές. Σε συνδυασμό με τη χαμηλή συγκέντρωση στην αγορά, αυτό σημαίνει ότι οι παραγωγοί αντιμετωπίζουν σε μεγάλο βαθμό τον ανταγωνισμό σε επίπεδο κόστους και τιμών. Αυτός ο ανταγωνισμός για το κόστος αντικατοπτρίζεται στη στρατηγική των παραγωγών, με πολλούς παραγωγούς να προβαίνουν σε offshoring και outsourcing παραγωγές στην Κίνα για να μειώσουν το κόστος. Στην παραγωγή παιχνιδιών, τα περιθώρια σε ολόκληρο τον κλάδο υφίστανται πιέσεις, με μακροπρόθεσμα περιθώρια κέρδους γύρω στο 6% για τις 100 πρώτες εταιρείες όσον αφορά το μέγεθος. Τα περιθώρια είναι χαμηλότερα για τις μικρές και μεσαίες επιχειρήσεις (ΜΜΕ) από ό, τι για τις μεγάλες επιχειρήσεις. Επίσης, το περιθώριο κέρδους για το λιανικό εμπόριο είναι χαμηλότερο από ό, τι για την κατασκευή παιχνιδιών.

Ο σύντομος κύκλος ζωής των παιχνιδιών οδηγεί σε εμφανή ανάγκη για καινοτομία, έρευνα και ανάπτυξη (R&D). Η καινοτομία αναγνωρίζεται ευρέως στον κλάδο ως απαραίτητη προϋπόθεση για τη διατήρηση μιας ανταγωνιστικής θέσης. Επίσης, οι στρατηγικές μάρκετινγκ είναι πολύ σημαντικές στον τομέα των παιχνιδιών. Το κλειδί για επιτυχημένες αποφάσεις στους τομείς του μάρκετινγκ αλλά και της έρευνας για νέα προϊόντα είναι η έρευνα αγοράς και η εισαγωγή καινοτομιών.

Ο ανταγωνισμός σε επίπεδο τιμών και καινοτομίας είναι πιθανόν να παραμείνουν έντονα σε μια δυναμική αγορά όπως αυτή των παιχνιδίων. Παρά τις γενικά θετικές προβλέψεις ανάπτυξης, τα παραδοσιακά και τα συμβατικά παιχνίδια θα αντιμετωπίσουν αυξημένο ανταγωνισμό από τα βιντεοπαιχνίδια και την πρόσφατη τάση στη χρήση των “tablets” και των έξυπνων τηλεφώνων για λόγους ψυχαγωγίας.

Πολλές τάσεις ενισχύουν τις προοπτικές αύξησης του ανταγωνισμού για παραδοσιακά παιχνίδια. Πρώτον, ο αριθμός των παιδιών μεταξύ 0-14 ετών στις ώριμες αγορές είναι πιθανό να σταθεροποιηθεί ή να μειωθεί στο εγγύς μέλλον. Δεύτερον, καθώς τα παιδιά ωριμάζουν σε μικρότερη ηλικία, η περίοδος παιχνιδιού θα είναι μικρότερη. Ως εκ τούτου, οι παραγωγοί θα αντιμετωπίσουν μεγαλύτερο ανταγωνισμό από υποκατάστατα παραδοσιακών και συμβατικών παιχνιδιών όπως τα βιντεοπαιχνίδια, τα "tablets" και τα έξυπνα τηλέφωνα που τείνουν να κυριαρχούν περισσότερο στις προτιμήσεις καθώς τα παιδιά ωριμάζουν.

Από τη θετική πλευρά, υπάρχουν εξελίξεις που δικαιολογούν τη συνεχή εξέλιξη της ανάπτυξης για την αγορά παραδοσιακών και συμβατικών παιχνιδιών. Πρώτον, η αγοραστική δύναμη αυξάνεται στις αναδυόμενες αγορές. Μια δεύτερη εξέλιξη με μεγάλες δυνατότητες για τα παραδοσιακά και τα συμβατικά παιχνίδια είναι η άνοδος των παιχνιδιών που επιτρέπουν την αναπαραγωγή παραδοσιακών παιχνιδιών σε ηλεκτρονικές πλατφόρμες και την αλληλεπίδραση μεταξύ φυσικών παιχνιδιών και εφαρμογές σε "tablets" και έξυπνα τηλέφωνα. Αρκετά παραδείγματα δείχνουν ότι οι παραγωγοί παιχνιδιών της ΕΕ εισέρχονται σε αυτές τις νέες πλατφόρμες, συχνά σε συνεργασία με τη βιομηχανία ψηφιακής ψυχαγωγίας. Για να συμβαδίσουν με την αγορά αυτή και να βελτιστοποιήσουν τις δυνατότητές της για παραδοσιακά και συμβατικά παιχνίδια, οι προμηθευτές παιχνιδιών της ΕΕ θα πρέπει να συμβαδίζουν με τους ανταγωνιστές των ΗΠΑ και της Ασίας σε αυτόν τον ταχέως αναπτυσσόμενο τομέα.

Όσον αφορά την τμηματοποίηση των προϊόντων, τα παιχνίδια κατασκευής και τα υπαίθρια αθλητικά παιχνίδια δείχνουν την υψηλότερη πρόβλεψη ανάπτυξης μεταξύ των παραδοσιακών παιχνιδιών. Τα επιτραπέζια παιχνίδια και παζλ δείχνουν σταθερές προβλέψεις για το μερίδιο αγοράς τους, καθώς αντιμετωπίζουν τον πιο άμεσο ανταγωνισμό από τα βιντεοπαιχνίδια, τα tablet και τις εφαρμογές έξυπνων τηλεφώνων.

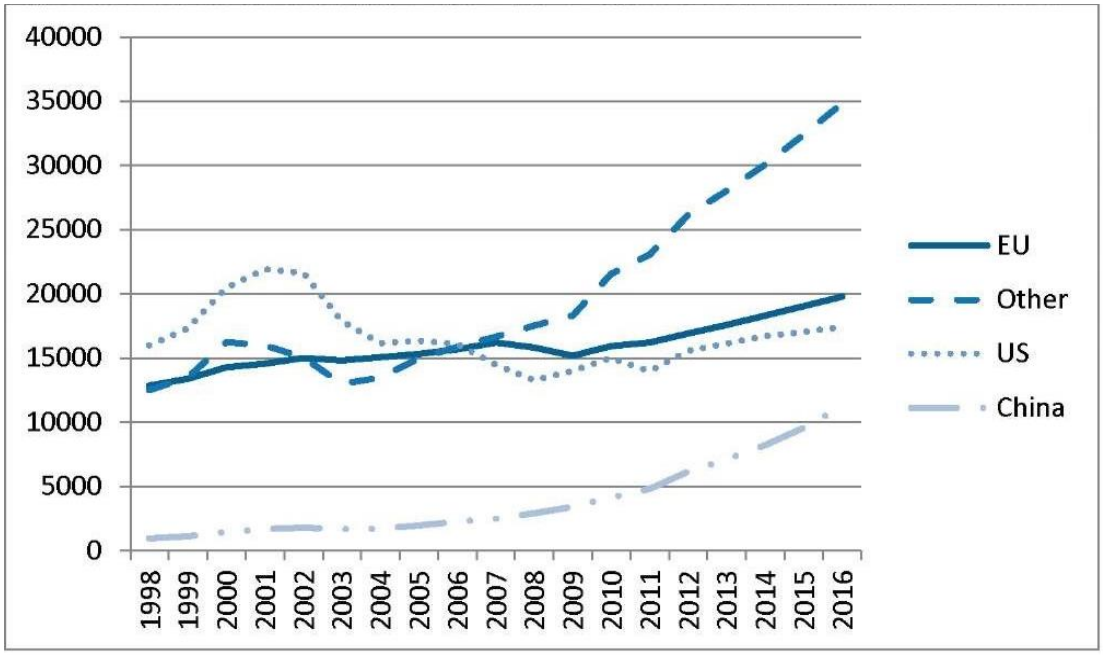
Τα πιστοποιημένα παιχνίδια παραμένουν μια μεγάλη και σταθερή πηγή του κύκλου εργασιών στην παραδοσιακή βιομηχανία παιχνιδιών. Είναι μοντέρνα αλλά επίσης προσφέρουν σταθερή ζήτηση και μειώνουν τον κίνδυνο επιτυχούς τοποθέτησης νέων προϊόντων στην αγορά, λόγω της σύνδεσης με την καθιερωμένη βιομηχανία ψυχαγωγίας. Επιπλέον, οι γονείς συνδέουν τα πιστοποιημένα παιχνίδια με καθιερωμένα εμπορικά σήματα που έχουν εγγυημένα υψηλή ποιότητα και ασφάλεια.

Η κύρια τάση λιανικής πώλησης παραδοσιακών και συμβατικών παιχνιδιών σε ολόκληρη την ΕΕ είναι η άνοδος του διαδικτυακού καναλιού λιανικής πώλησης. Οι πωλήσεις μέσω διαδικτύου παρουσιάζουν διψήφιους ρυθμούς ανάπτυξης και έχουν ήδη αποκτήσει μερίδια αγοράς σχεδόν 20% σε ορισμένες ώριμες αγορές. Οι αγορές της Νότιας Ευρώπης ήταν λιγότερο προσανατολισμένες στις ηλεκτρονικές αγορές, αλλά και αυτό φαίνεται ότι μεταβάλλεται ταχύτατα.

4.1.2. Χαρακτηριστικά του κλάδου

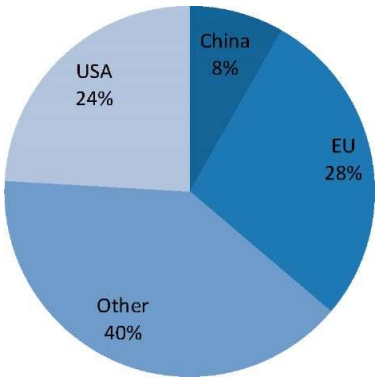
4.1.2.1. Κύρια προϊόντα και κατασκευαστές

Οι συνολικές παγκόσμιες πωλήσεις παραδοσιακών και συμβατικών παιχνιδιών ανήλθαν το 2011 στα 58 δισ. €. Το Ηνωμένο Βασίλειο, η Γαλλία, η Γερμανία, η Ιταλία και η Ισπανία είναι οι μεγαλύτερες αγορές παιχνιδιών στην ΕΕ. Ενώ η αγορά των παραδοσιακών και συμβατικών παιχνιδιών παρουσίασε ανάπτυξη στις αρχές της δεκαετίας του 2000, η ανάπτυξη σημείωσε πάλι πτώση μετά από λίγα χρόνια (Εικόνα 19). Η αμερικανική αγορά είναι η εξαίρεση στον κανόνα αυτού, καθώς δεν επέστρεψε στην κορύφωσή των ετών 2001-2002. Οι αγορές της ΕΕ και των ΗΠΑ παρέμειναν στάσιμες κατά την πρόσφατη οικονομική και χρηματοπιστωτική κρίση στα έτη 2008-2011. Η κρίση στην ΕΕ υποχώρησε νωρίς και σημειώθηκε κάποια ανάκαμψη μετά το 2009, ενώ η αγορά των ΗΠΑ μειώθηκε το 2011. Τα στοιχεία που παρουσιάζονται παρακάτω είναι προβλέψεις από το 2012 και μετά.



Εικόνα 20: Οι πωλήσεις παραδοσιακών παοχνιδιών σε εκ. € (πρόβλεψη για το 2012-2016). Πηγή: Euromonitor

Παρά το γεγονός ότι οι πωλήσεις αυξάνονταν με την πάροδο του χρόνου, το μερίδιο αγοράς της ΕΕ στην παγκόσμια αγορά παιχνιδιών παρέμεινε σταθερό, ενώ οι ΗΠΑ έχασαν μερίδιο αγοράς μετά το 2001. Η αγορά της ΕΕ αντιπροσωπεύει σήμερα το 28% των παγκόσμιων παραδοσιακών πωλήσεων παιχνιδιών (Εικονα 21).



Εικόνα 21: Οι πωλήσεις συμβατικών παιχνιδιών ανά χώρα, 2011. Πηγές: Euromonitor, Ecorys

Τα παραδοσιακά παιχνίδια κυριαρχούνται από παγκόσμιες μάρκες, όπως η LEGO, η Hasbro και η Mattel. Στις περισσότερες μεγάλες χώρες της ΕΕ οι τρεις αυτές εταιρείες κατέχουν ηγετική θέση στην αγορά, με το συνολικό μερίδιο αγοράς τους στην ΕΕ να είναι σχεδόν 27%.

Πίνακας 3: Οι 10 μεγαλύτερες επιχειρήσεις του κλάδου στην Ε.Ε. Πηγή: Euromonitor

Εταιρεία	εκ. €	% του συνόλου
Mattel Inc	1343.6	10.08
LEGO Group	1108.5	8.32
Hasbro Inc	1084.9	8.14
Private Label	575.2	4.32
Simba-Dickie Group GmbH & Co KG	386.2	2.90
Giochi Preziosi SpA	375.7	2.82
Geobra Brandstatter GmbH & Co KG	316.8	2.38
VTech Holdings Ltd	296.9	2.23
Ravensburger AG	234.1	1.76
Takara Tomy Co Ltd	203.6	1.53
Other	7403.9	55.55

Δίπλα στους παγκόσμιους παίκτες υπάρχουν και ορισμένοι τοπικοί. Στην Ιταλία υπάρχει ο ισχυρός παίκτης Giochi Preziosi Spa, ο οποίος είναι επίσης παρών στη Γαλλία. Μια άλλη ισχυρή εγχώρια ιταλική εταιρεία, Clementoni Spa, είναι πολύ δραστήρια στην Ιταλία αλλά όχι έξω από τη χώρα. Μια τέτοια διακύμανση μπορεί επίσης να παρατηρηθεί κατά τη σύγκριση της παγκόσμιας και της κοινοτικής αγοράς. Σημαντικός παίκτης στον κόσμο, όσον αφορά τις πωλήσεις, ο όμιλος BANDAI NAMCO, δεν είναι καν στην πρώτη θέση στην ΕΕ. Μια άλλη κορυφαία εταιρεία και 5^{η} παγκοσμίως, η Takara Tomy Co Ltd, εισέρχεται μόλις στην πρώτη δεκάδα στην αγορά της ΕΕ, κυρίως λόγω της παρουσίας της στην αγορά του Ηνωμένου Βασιλείου.

Οι κορυφαίες 5 εταιρείες στις ΗΠΑ έχουν πάνω από το 50% της αγοράς. Στις ΗΠΑ, η Mattel Inc. ήταν ο μεγαλύτερος κατασκευαστής το 2011 με πωλήσεις αξίας περίπου 3,3 δισ. ευρώ, αντιπροσωπεύοντας το 23,5% της παραδοσιακής αγοράς παιχνιδιών. Η Hasbro κατέχει τη δεύτερη θέση με περίπου 17% της αγοράς, ακολουθούμενη από τη LEGO στην τρίτη θέση.

Στην Κίνα, κάθε μία από τις 5 πρώτες εταιρείες αντιπροσωπεύει μεμονωμένα λιγότερο από το 4% της αγοράς και στο σύνολο τους αντιπροσωπεύουν λιγότερο από το 10% της συνολικής αγοράς όσον αφορά τις πωλήσεις. Guangdong Alpha Animation & Culture Co Ltd είναι ένας σημαντικός εγχώριος παίκτης στην Κίνα. Αντιπροσωπεύει το 3,2 % της κινεζικής αγοράς παραδοσιακών παιχνιδιών. Η Guangdong Alpha επικεντρώνεται κυρίως στην εγχώρια αγορά. Επιπλέον, εξάγει παιχνίδια στο Ηνωμένο Βασίλειο και σε ορισμένες περιφερειακές και γειτονικές αγορές.

Η Mattel, η Hasbro και η LEGO είναι ηγέτες στην αγορά παγκοσμίως για παραδοσιακά και συμβατικά παιχνίδια. Ενώ η Mattel είναι η κορυφαία εταιρεία στον κόσμο από την άποψη των πωλήσεων, στην Κίνα κατέχει την 4η θέση με μερίδιο 0,8%. Ομοίως, η Hasbro αντιπροσωπεύει περίπου το 0,8% της κινεζικής αγοράς, αλλά δεν κατατάσσεται στις κορυφαίες 5 εταιρείες όσον αφορά την αξία των πωλήσεων στην Κίνα.

4.1.2.2. Υπηρεσίες λιανικής πώλησης παιχνιδιών

Το 2010, η λιανική πώληση παιχνιδιών στον τομέα των εξειδικευμένων καταστημάτων στην ΕΕ (27 χώρες) συνίστατο σε 19.129 επιχειρήσεις που απασχολούσαν συνολικά 101.800 εργαζόμενους.

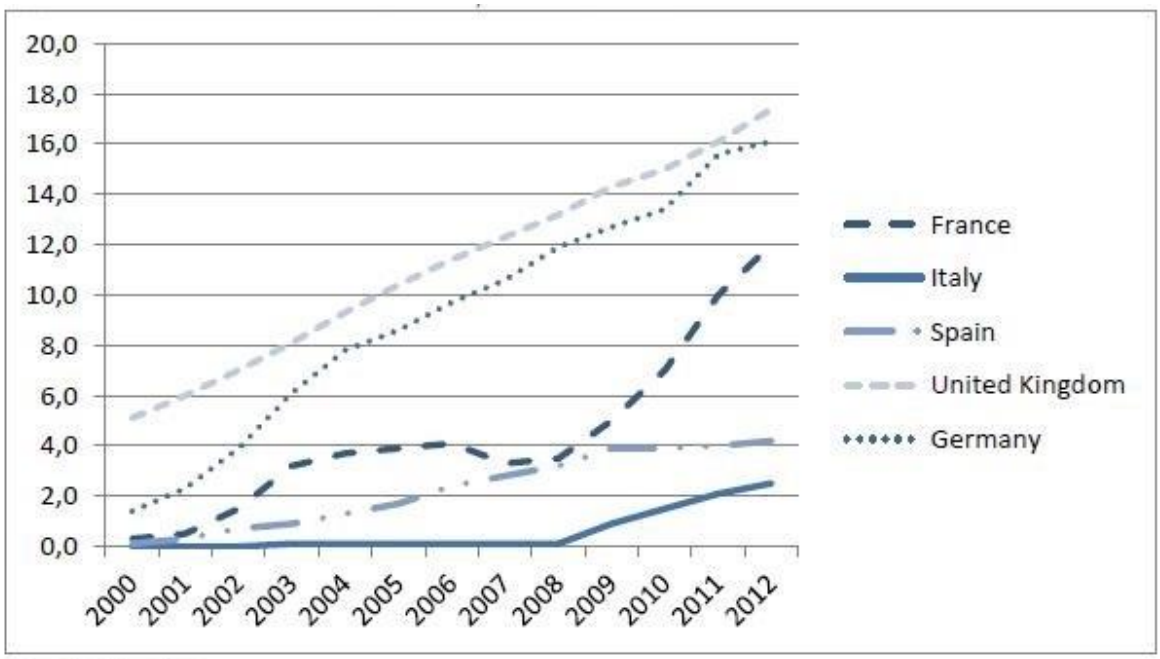
Ο πίνακας που ακολουθεί παρουσιάζει τα κύριως δίκτυα λιανικής πώλησης στις κυριότερες ευρωπαϊκές αγορές παιχνιδιών για το 2011. Όπως φαίνεται από τον πίνακα, τα περισσότερα παιχνίδια πωλούνται σε παραδοσιακά καταστήματα λιανικής πώλησης. Ωστόσο, οι πωλήσεις παιχνιδιών στο διαδίκτυο αυξάνονται ραγδαία, αυξάνοντας τα μερίδια αγοράς σε πολλές χώρες. Στην Ευρώπη τα κυριότερα κανάλια διανομής παραμένουν τα εξειδικευμένα καταστήματα παιχνιδιών, ενώ η δεύτερη θέση λαμβάνεται από τους λοιπούς λιανοπωλητές.

Πίνακας 4: Το μερίδιο αγοράς των δικτύων λιανικής πώλησης παιχνιδιών στην Ευρώπη

	Ηνωμένο Βασίλειο	Γαλλία	Γερμανία	Ιταλία	Ισπανία
Grocery retailers	28,8%	40,7%	12,4%	33,0%	23,1%
Electronics and Appliance specialist retailers	2,2%	n/a	1,0%	n/a	n/a
Mixed retailers	13,4%	1,3%	11,7%	7,0%	25,5%
Leisure and Personal Goods specialist retailers	37,6%	46,5%	46,1%	57,8%	44,0%
- Traditional toys and games store	28,6%	45,4%	41,2%	50,4%	40,0%
- Media product stores	1,8%	n/a	0,6%	1,3%	n/a
- Others	7,2%	1,1%	4,3%	6,1%	4%
Other non-grocery retailers	0,5%	0,6%	6,7%	n/a	2,7%
Vending	0,1%	0,1%	0,4%	0%	n/a
Homeshopping	0,8%	0,1%	1,4%	0,1%	0,4%
Internet retailing	16,1%	10%	15,6%	2,1%	4,0%
Direct selling	0,4%	0,1%	3,5%	0%	0,3%
Other	0.1%	0.6%	1.2%	0%	0%
Total	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**

Αν και τα περισσότερα παιχνίδια πωλούνται σε παραδοσιακά καταστήματα λιανικής πώλησης, τα τελευταία χρόνια οι ηλεκτρονικές πωλήσεις έχουν προσελκύσει αρκετή προσοχή. Οι περισσότερες αλυσίδες και κατασκευαστές λιανικής πώλησης άρχισαν να επενδύουν σε διαδικτυακές λιανικές πωλήσεις (e-tailing) και για τολόγο αυτό διαθέτουν ιστοσελίδες για πωλήσεις (*Πηγή: Euromonitor International, 2012*). Επιπλέον, οι κατασκευαστές πωλούν επίσης μέσω εξειδικευμένων διαδικτυακών εμπόρων όπως το Amazon. Τέτοια e-tailing δίνουν επίσης τη δυνατότητα στους μικρούς κατασκευαστές να βρουν το δρόμο τους στην αγορά των παιχνιδιών. Ενώ οι τιμές σε ηλεκτρονικά καταστήματα τείνουν να είναι χαμηλότερες από ό, τι στα παραδοσιακά καταστήματα (*Πηγή: Euromonitor International, 2012*), αυτό δεν συνεπάγεται πάντα μικρότερο περιθώριο για τους παραγωγούς και τον ηλεκτρονικό έμπορο λιανικής. Η εξοικονόμηση του κόστους εφοδιαστικής αλυσίδας συνεπάγεται μεγαλύτερο περιθώριο κέρδους.

Το διάγραμμα που ακολουθεί (Εικόνα 21) παρουσιάζει τον τρόπο με τον οποίο το λιανικό εμπόριο στο Διαδίκτυο αύξησε το μερίδιό του με την πάροδο του χρόνου στις πέντε κύριες αγορές.



Εικόνα 22: Ποσοστό των αγορών μέσω διαδικτύου για τις 5 κύριες οικονομίες της Ε.Ε. Πηγή: Euromonitor

4.1.3. Καινοτομία, έρευνα και διαφήμιση

Λόγω των σχετικά μικρών κύκλων ζωής των προϊόντων και του ανταγωνισμού στις προτιμήσεις των παιδιών με άλλα προϊόντα, η καινοτομία αποτελεί σημαντικό στοιχείο των επιχειρηματικών μοντέλων των κατασκευαστών παιχνιδιών. Οι καινοτομίες μπορούν να αποτελούνται από εντελώς νέα παιχνίδια που περιλαμβάνουν μια νέα έννοια ή λειτουργικότητα, όπως τα διαδραστικά παιχνίδια. Εναλλακτικά, πρέπει να αφορούν νέα θέματα και ενημερωμένες έννοιες, σε επιτραπέζια παιχνίδια και σε συστήματα παιχνιδιών όπως τα LEGO ή τα Playmobil. Η καινοτομία αντανακλά την επιτυχή ανάπτυξη και πορεία των προϊόντων, η οποία εξαρτάται τόσο από τις επενδύσεις στη διαφήμιση όσο και από την έρευνα αγοράς και τεχνολογίας.

Παρόλο που δεν υπάρχουν διαθέσιμα δημόσια στοιχεία σχετικά με τις επενδύσεις στο μάρκετινγκ των εταιρειών παιχνιδιών, οι πληροφορίες από παραγωγούς και ενώσεις κατασκευαστών δείχνουν ότι οι επενδύσεις είναι μικρές σε σχέση με τον κύκλο εργασιών. Οι διαφημιστές επικεντρώνονται σε τηλεοπτικές διαφημίσεις, καταλόγους προϊόντων και όλο και περισσότερο στα κοινωνικά μέσα. Οι προμηθευτές και οι λιανοπωλητές παιχνιδιών μικρών και μεσαίων επιχειρήσεων δεν διαθέτουν τα οικονομικά μέσα για την πραγματοποίηση σημαντικής τηλεοπτικής διαφήμισης. Τα διαδικτυακά κανάλια πωλήσεων, όπως προαναφέρθηκε, προσφέρουν νέους τρόπους διαφήμισης για τις ΜΜΕ για να αυξήσουν την προβολή τους, όπως η ανάπτυξη ιστότοπων, η συμμετοχή σε κοινωνικά μέσα και η πρόσβαση σε εξειδικευμένους διαδικτυακούς εμπόρους λιανικής πώλησης.

4.1.4. Προβλέψεις της αγοράς

Η βιομηχανία παιχνιδιών είναι πολύ δυναμική και χαρακτηρίζεται από υψηλό ανταγωνισμό στην καινοτομία και στις τιμές πώλησης. Τα τελευταία χρόνια η παραδοσιακή και συμβατική βιομηχανία παιχνιδιών γνώρισε αυξανόμενο ανταγωνισμό από τα βιντεοπαιχνίδια, τα οποία και θα παραμείνουν ως η μεγαλύτερη πρόκληση για την ανάπτυξή της (*πηγή: Euromonitor International, 2012*). Επίσης, μεγάλη πίεση στην παραδοσιακή συμβατική βιομηχανία παιχνιδιών ασκεί η αυξανόμενη ζήτηση των καταναλωτών για παιχνίδια με ηλεκτρονικά εξαρτήματα. Μια παρόμοια τάση είναι και η αύξηση των "tablets" και των έξυπνων τηλεφώνων, τα οποία λειτουργούν ανταγωνιστικά στις προτιμήσεις των παιδιών και των εφήβων.

Η βιομηχανία παιχνιδιών αντιμετωπίζει ορισμένες εξελίξεις που ενδέχεται να περιορίσουν τις προοπτικές ανάπτυξης της αγοράς, όπως:

- Ο αριθμός των παιδιών ήταν σταθερός στις ΗΠΑ και δείχνει μια πτωτική τάση στην ΕΕ από τις αρχές της δεκαετίας του 2000.
- Τα παιδιά ωριμάζουν νωρίτερα, γεγονός που υποδηλώνει ότι η περίοδος παιχνιδιού είναι μικρότερη, γεγονός που θα επηρεάσει αρνητικά την πιθανή αγορά (πηγή: Van Lotringen International, 2005).
- Τα παιδιά μεταβαίνουν σε νεαρή ηλικία από παραδοσιακά και συμβατικά παιχνίδια σε ηλεκτρονικά παιχνίδια και βιντεοπαιχνίδια.

Ένας εξωτερικός παράγοντας που επηρεάζει θετικά τις προοπτικές για την παραδοσιακή αγορά παιχνιδιών είναι ότι η αγοραστική δύναμη ανά παιδί αυξάνεται, ειδικά στις αναδυόμενες αγορές όπως η Κίνα.

Όσον αφορά την τμηματοποίηση της αγοράς, τα παιχνίδια κατασκευής έχουν επιδείξει τη μεγαλύτερη ανάπτυξη και αναμένεται να συνεχίσουν να το κάνουν. Αυτό μπορεί να σχετίζεται με τη σημασία της αξίας του παιχνιδιού για τους γονείς. Οι γονείς εξακολουθούν να αποτελούν σημαντικό παράγοντα στην επιλογή των παιχνιδιών. Τα παιχνίδια κατασκευής θεωρούνται ότι διεγείρουν τη δημιουργικότητα και μπορούν να χρησιμοποιηθούν διαφορετικά κάθε φορά που ένα παιδί παίζει μαζί τους. Επίσης, τα υπαίθρια και αθλητικά παιχνίδια αναμένεται να παρουσιάσουν μεγάλη ανάπτυξη λόγω των αυξανόμενων ανησυχιών των γονέων για την παιδική παχυσαρκία και άλλα προβλήματα υγείας. Ταυτόχρονα, τα εσωτερικά παιχνίδια όπως τα επιτραπέζια παιχνίδια και τα παζλ τείνουν να έχουν μάλλον σταθερό μερίδιο αγοράς. Ως ανταπόκριση στον ανταγωνισμό και στη μετατόπιση της ζήτησης προς ηλεκτρονικά παιχνίδια, οι παραδοσιακοί κατασκευαστές παιχνιδιών άρχισαν να διερευνούν τις δυνατότητες δημιουργίας παιχνιδιών που συνδέονται με ηλεκτρονικά gadgets (*πηγή: Euromonitor International, 2012*).

Η τρέχουσα εξέλιξη στην αγορά παιχνιδιών ασκεί μεγαλύτερη πίεση στη βιομηχανία παιχνιδιών για να αναζητήσει διαφορετικούς κερδοφόρους τρόπους για τη διαχείριση της αβεβαιότητας της ζήτησης μέσω του μάρκετινγκ, της αδειοδότησης και της καινοτομίας. Η μετατόπιση στην τεχνολογία των "tablets" και η ασύρματη επικοινωνία που είναι εμφανής στα καταναλωτικά αγαθά δίνουν κίνητρα στους παραδοσιακούς κατασκευαστές παιχνιδιών να διερευνήσουν το συνδυασμό παραδοσιακών παιχνιδιών με ηλεκτρονικά gadgets. Για παράδειγμα, η Mattel προσπαθεί να κερδίσει κάποια δύναμη στην αγορά των βιντεοπαιχνιδιών αποκτώντας παιχνίδια Radica ((*πηγή: Euromonitor International, 2012*).

Το 2012, η Mattel εγκαινίασε την Apptivity, μια ενεργή τεχνολογία αφής που επιτρέπει στα φυσικά παιχνίδια να αλληλεπιδρούν με το iPad. Ένας από τους βασικούς ανταγωνιστές της Mattel, Hasbro,

έχει επίσης ευθυγραμμιστεί με την Electronic Arts προκειμένου να αναβαθμίσει τα παραδοσιακά παιχνίδια του, όπως το Monopoly σε βιντεοπαιχνίδια (Euromonitor International, 2012a). Οι ευρωπαίοι παραγωγοί επιτραπέζιων παιχνιδιών παζλ, όπως οι Ravensburger και Jumbo, προσφέρουν επίσης παραδοσιακά παιχνίδια σε ηλεκτρονικές πλατφόρμες ή προσθέτουν ηλεκτρονικά αξεσουάρ στα παραδοσιακά προϊόντα τους. Ως τελευταίο παράδειγμα, η Lego δημιούργησε μια εφαρμογή iPhone που προτρέπει τα παιδιά να κατασκευάσουν γρήγορα μικρά μοντέλα με φυσικά μπλοκ. Ως εκ τούτου, τα παιχνίδια cross-over, τα οποία επιτρέπουν στα φυσικά παιχνίδια να αλληλεπιδρούν με τις τεχνολογίες, αποτελούν μία από τις θέσεις που πρόσφατα έλαβαν την προσοχή των παραδοσιακών κατασκευαστών παιχνιδιών και οι οποίες ενδέχεται να επιτρέψουν στους παραδοσιακούς κατασκευαστές παιχνιδιών να δημιουργήσουν νέες αγορές και να κάνουν τη ζήτηση πιο σταθερή.

4.2. Έρευνα Αγοράς στο διαδίκτυο και τα κοινωνικά μέσα

Με βάση την ανάλυση του κλάδου που έχει προηγηθεί, εύλογα συμπεραίνει κανείς ότι η βιομηχανία των συμβατικών παιχνιδιών είναι ένας κλάδος σχετικά σταθερός στην πορεία του χρόνου ως προς τα χαρακτηριστικά του, ο οποίος όμως έχει να αντιμετωπίσει μια σειρά από σύγχρονες προκλήσεις. Επομένως, η ανάγκη για έρευνα και επένδυση στην κατεύθυνση των νέων τάσεων της αγοράς αλλά και για σύμπλευση με το ενδιαφέρον των καταναλωτών, κρίνεται επιτακτική.

Για να διατηρήσουν την ανταγωνιστικότητά τους στα νέα δεδομένα, πρωτίστως οι κατασκευαστές και δευτερευόντως οι έμποροι παραδοσιακών και συμβατικών παιχνιδιών πρέπει να δημιουργήσουν μια σύγχρονη εικόνα με δυναμική παρουσία στο διαδίκτυο και τα κοινωνικά μέσα, τόσο για την προώθηση των προϊόντων τους όσο και για να μπορούν να αφουγκραστούν καλύτερα τις επιθυμίες των καταναλωτών. Μπορούμε να συνοψίσουμε τα συμπεράσματα των κεφαλαίων 1 και 4 ως δράσεις τις οποίες θα πρέπει να ακολουθήσει μια εταιρεία του κλάδου της κατασκευής παιχνιδιών και που αφορούν τα κανάλια του διαδικτύου ως εξής:

1. Δημιουργία και διατήρηση δυναμικής παρουσίας στο διαδίκτυο και τα κοινωνικά μέσα, με έμφαση στην ενθάρρυνση της αλληλεπίδρασης των χρηστών
2. Διερεύνηση των προτιμήσεων του καταναλωτικού κοινού και των τάσεων της αγοράς σε πραγματικό χρόνο με εκμετάλλευση των καναλιών/πηγών του διαδικτύου
3. Δημιουργία και συντήρηση των κατάλληλων διαφημιστικών εκστρατειών στα επιμέρους κανάλια για επιτυχημένη προώθηση των προϊόντων
4. Παρακολούθηση των αντιδράσεων των χρηστών για τα προϊόντα (κριτικές, σχόλια) μέσα από τα επιμέρους κανάλια, και ανατροφοδότηση της διαδικασίας

Για τους σκοπούς της παρούσας διπλωματικής, που περιορίζονται στη 2η δράση *(Διερεύνηση των προτιμήσεων του καταναλωτικού κοινού και των τάσεων της αγοράς)* θα χρησιμοποιήσουμε την υπόθεση ότι τα κανάλια στα οποία δραστηριοποιούνται οι χρήστες και ο κλάδος στο διαδίκτυο είναι ενιαία. Δηλαδή, ότι τα ίδια σύνολα δεδομένων στα οποία εκφράζονται οι προτιμήσεις και άρα οι τάσεις της αγοράς, είναι αυτά που συγκεντρώνουν και την αλληλεπίδραση των χρηστών σε σχέση με τα προϊόντα και τη διαφήμιση. Η υπόθεση αυτή είναι λογική και ασφαλής, μιας και οι δραστηριότητες είναι αλληλοεξαρτώμενες, και η εμπειρία των χρηστών ενιαία, ανεξάρτητα από τους σκοπούς της επιχείρησης που εξυπηρετώνται κάθε φορά.

4.2.1. Η παρουσία του κλάδου στα κοινωνικά μέσα

Ο ψηφιακός κόσμος κυριαρχεί σε ένα σημαντικό ποσοστό στη διάρκεια ζωής του μέσου καταναλωτή – ένα 80% των καταναλωτών διερευνούν τις επιλογές τους στο διαδίκτυο πριν πραγματοποιήσουν κάποια αγορά. Για να ελεγχθεί αυτή η καταναλωτική συμπεριφορά, οι εταιρείες πρέπει να εντάξουν στη λειτουργικότητά τους το μάρκετινγκ στα κοινωνικά μέσα (social media marketing). Αυτό ισχύει και για όσες δραστηριοποιούνται στον κλάδο της κατασκευής συμβατικών και παραδοσιακών παιχνιδιών. Κάποιοι από τους κατασκευαστές δεν έχουν συνειδητοποιήσει ακόμα την αξία της παρουσίας τους στα κοινωνικά μέσα, ενώ κάποιες εταιρείες έχουν δώσει ιδιαίτερη βάση, αναγνωρίζοντας ότι τα ψηφιακά κανάλια παράγουν σημαντικά και μετρήσιμα αποτελέσματα προς όφελός τους. Η Brand Chorus δημιούργησε μια παγκόσμια κατάταξη για το κατά πόσον οι εταιρείες του κλάδου δραστηριοποιούνται στα δημοφιλέστερα κοινωνικά μέσα Twitter, Facebook, YouTube και Instagram (Πηγή: “The best toy brands on social media,” telegraph.co.uk, 2017).

Ενδεικτικά:

- Η Hot Wheels έχει σχεδόν 1.4 εκ.ακολούθους σε τέσσερις πλατφόρμες κοινωνικών δικτύων, με κορυφαία την παρουσία της στο Twitter και στο Instagram. Το περιεχόμενο αυτό όμως ήταν στοχευμένο περισσότερο προς το ενήλικο κοινό (συλλέκτες) παρά στα παιδιά ή τους γονείς.

- Η Melissa & Doug, γνωστή για τα εκπαιδευτικά παιχνίδια και τα ξύλινα παζλ, είναι μία από τις πιο δραστήριες εταιρείες στα κοινωνικά μέσα, και παρόλο που δεν έχει παρουσία στο Instagram και το Youtube έρχεται 9η στην κατάταξη.

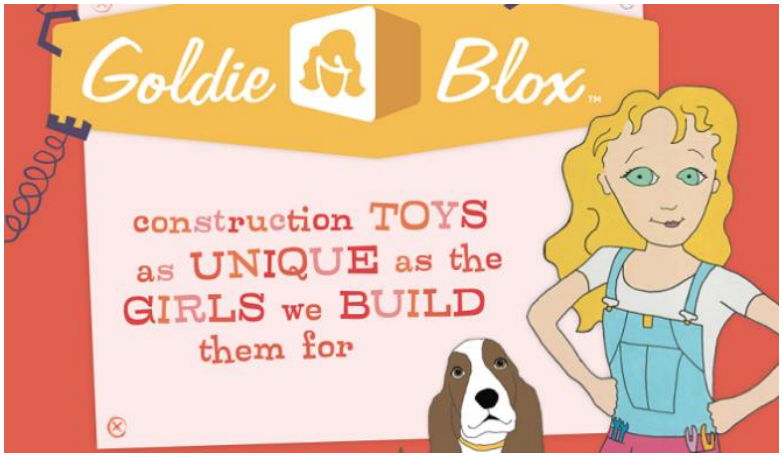
- Η Fisher-Price δημοσιεύει αναρτήσεις συστηματικά και εξίσου σε όλα τα μέσα κοινωνικής δικτύωσης, προωθώντας μάλιστα και τη σειρά βίντεο #LetsTalkToys όπου τα ίδια τα παιδιά αξιολογούν τα παιχνίδια της, δημιουργώντας σημαντικό αριθμό αλυσίδων αναδημοσιεύσεων.

- Η Barbie, παρόλο που έχει το μεγαλύτερο απόλυτο αριθμό ακολούθων στα κοινωνικά μέσα, δημοσιεύει μόνο 1-2 φορές την ημέρα, δημιουργώντας όμως αντίκτυπο λόγω της δημοφιλίας της. Έτσι κατατάσσεται στην τέταρτη θέση της λίστας.

- Η GoldieBlox, η εταιρεία που δημιουργεί εκπαιδευτικά παιχνίδια με έμφαση στη μηχανική και την επίλυση προβλημάτων για κορίτσια, έχει προσπεράσει στο σκορ εδραιωμένες εταιρείες όπως η Barbie, η Fisher-Price και η Hot Wheels. Έχει περίπου μισούς ακολούθους από ότι οι προαναφερθείσες, αλλά έχει έντονη αλληλεπίδραση με τους χρήστες, λόγω των ισχυρών θεματικών της σε σχέση με την ενδυνάμωση των κοριτσιών και την εξάλειψη των φυλετικών διακρίσεων στα παιδικά παιχνίδια.

- Η Monster High, που έχει λανσαριστεί από τη Mattel βρίσκεται στη δεύτερη θέση, με το περισσότερο δημιουργημένο από χρήστες περιεχόμενο και την περισσότερη χρήση του Instagram σε σχέση με τους ανταγωνιστές της.

- Την πρώτη θέση έχει η Lego, η οποία έχει γίνει πλέον ο μεγαλύτερος κατασκευαστής παιχνιδιών μετά από την έκδοση της ταινίας «The Lego Movie». Η Lego έχει περισσότερα βίντεο στο YouTube, συμπεριλαμβανόμενης μιας σειράς βίντεο «Κάν’το μόνος σου» ("How To" videos).



Εικόνα 23: Η Goldie Blox, μια εταιρεία με μεγάλη αλληλεπίδραση χρηστών στα κοινωνικά μέσα

4.2.2. Προώθηση και ερευνα αγοράς στα κανάλια του διαδικτύου

4.2.2.1. Προώθηση και ερευνα αγοράς στα κοινωνικά μέσα

Πολλές εταιρείες κατασκευής παιχνιδιών είναι αναμενόμενο να θέλουν να δώσουν έμφαση στη δύναμη του οπτικού μάρκετινγκ. Οι οπτικές πλατφόρμες κοινωνικών δικτύων, όπως το Pinterest, το Snapchat και το Instagram δίνουν τη δυνατότητα στους χρήστες να δουν τα παιχνίδια στην πράξη και κάνουν εύκολη την αλληλεπίδραση, καθώς οι γονείς μπορούν να ανεβάσουν φωτογραφίες δικές τους ή των παιδιών τους και να σχολιάσουν αυτές των διαδικτυακών τους φίλων.

Οι πιο δημοφιλείς κοινωνικές πλατφόρμες Twitter και Facebook πρέπει επίσης να συμπεριληφθούν. Το Facebook παραμένει μακράν το μεγαλύτερο κοινωνικό δίκτυο, με 1.86 δις. χρήστες το μήνα. Η δομή του ενθαρρύνει την αλληλεπίδραση με διαφορετικούς τύπους περιεχομένου, και κάνει εύκολη τη δημιουργία κοινοτήτων και ομάδων γύρω από μια συγκεκριμένη θεματολογία (μάρκες, προϊόντα, κλάδος). Το Twitter μπορεί να χρησιμοποιηθεί επίσης για την προώθηση εικόνων, όπως επίσης για την έκφραση κριτικών και τη διευθέτηση ζητημάτων εξυπηρέτησης καταναλωτών.

Για να προσδιοριστεί η καλύτερη στρατηγική για κάποια συγκεκριμένη εταιρεία, οι κατασκευαστές οφείλουν να γνωρίζουν ποιο είναι το κοινό τους. Για παράδειγμα, τα δημογραφικά του Pinterest δείχνουν προς τις γυναίκες οι οποίες είναι γεννημένες τις δεκαετίες από 1980 έως 2000 (millennials). Μπορεί επομένως να αποτελέσει μια καλή επιλογή για πρόσβαση στο κοινό των νέων μαμάδων. Το Instagram και το Snapchat έχουν νεότερους χρήστες, και φαίνεται να είναι πιο κατάλληλα για την πρόσβαση στα παιδιά του γυμνασίου και του λυκείου.

Στην κατηγορία των χρηστών με μεγάλη επιρροή (influencers), οι συστηματικοί χρήστες του Youtube (YouTube video bloggers) μπορούν επίσης να αποδειχτούν χρήσιμοι σε εκστρατείες προώθησης στο διαδίκτυο, και να χτίσουν μια σχέση εμπιστοσύνης με το καταναλωτικό κοινό.

Ενώ για τους σκοπούς της προώθησης όλα τα κανάλια είναι εξίσου σημαντικά ανάλογα με το αγοραστικό κοινό και τις ιδιαιτερότητές του, η έρευνα τάσεων της αγοράς αξίζει να σημειωθεί ότι είναι σημαντικά δυσκολότερη στις πηγές που έχουν σαν κύριο περιεχόμενο αναρτήσεις πολυμέσων. Αυτό διότι απαιτεί την εμπλοκή προηγμένων αλγορίθμων επεξεργασίας εικόνας και βίντεο πάνω σε μεγάλα δεδομένα – ένας τομέας που δεν έχει διερευνηθεί επαρκώς ακόμα και πιθανότατα να

αποτελέσει το επόμενο τεχνολογικό ορόσημο για την Εξόρυξη Δεδομένων. Αντίθετα, και όπως έχει φανεί από το Κεφ.3 της παρούσας εργασίας, η πρόοδος που έχει γίνει στην Εξόρυξη Δεδομένων από αναρτήσεις κειμένου, αλλιώς Εξόρυξη Κειμένου (textual data mining ή text mining) είναι σημαντική και αναμένεται να εφαρμόζεται όλο και περισσότερο.

4.2.2.2. Η σημασία των κριτικών

Όπως έχει προαναφερθεί, η πλειοψηφία των μεγάλων αγορών και πολλές από τις μικρότερες αγορές, γίνονται αφού προηγηθεί κάποιου είδους έρευνα στο διαδίκτυο. Η εποχή όπου μια εταιρεία μπορούσε να παράγει και να προωθήσει επιτυχημένα με επιθετικό μάρκετινγκ προϊόντα χαμηλής ποιότητας, έχει περάσει ανεπιστρεπτί. Πλέον οι καταναλωτές έχουν ενεργό ρόλο στον κύκλο ζωής ενός προϊόντος από πολύ αρχικά στάδια, και η ανεμπόδιστη πρόσβαση στην πληροφορία τους επιτρέπει να επιλέξουν τις αγορές τους με αξιοπιστία και ασφάλεια.

Πριν την αγορά, ο σύγχρονος καταναλωτής επιδιώκει να εντοπίσει μέσα από κριτικές και αναφορές άλλων καταναλωτών ποιο είναι το καλύτερο προϊόν σε σχέση με τον ανταγωνισμό, αν κάποιο προϊόν είναι πιο δημοφιλές και αν όσοι το έχουν αγοράσει ήδη είναι ευχαριστημένοι από αυτό. Αυτό μπορεί να το επιτύχει μέσα από τις μηχανές αναζήτησης (Google), τα κοινωνικά μέσα, τα ηλεκτρονικά καταστήματα ή άλλες ιστοσελίδες όπως ιστολόγια (blogs) και ηλεκτρονικές κοινότητες (forums). Με τις ταχύτητες στις οποίες είναι όλες αυτές οι πηγές προσβάσιμες, είναι εύκολο για τον χρήστη να κάνει μια ολοκληρωμένη αναζήτηση, σε περισσότερες από μία πηγές τη φορά.

4.3. Πρόταση μοντέλου για ανίχνευση τάσεων στον κλάδο των παιχνιδιών

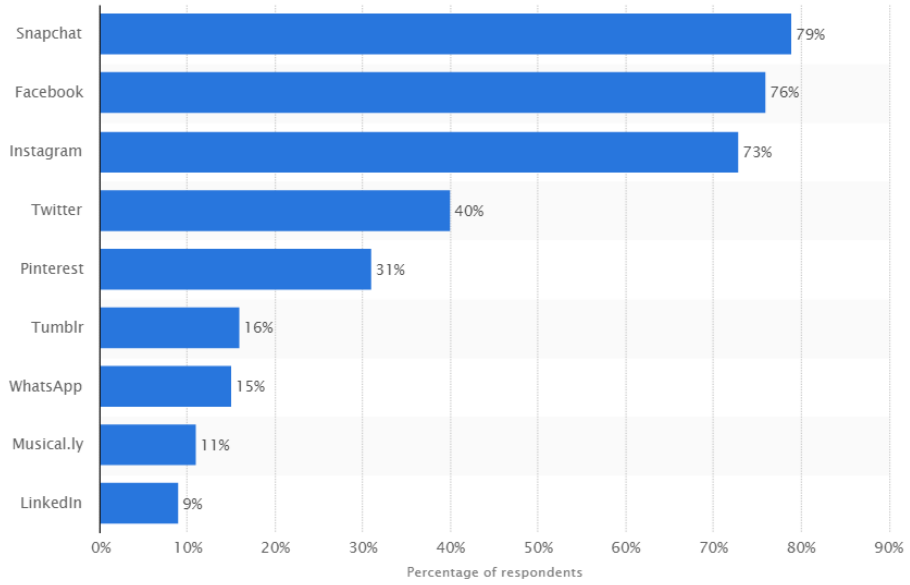
4.3.1. Επιλογή των πηγών δεδομένων

Προκειμένου να είναι εφικτή η πρόταση μιας ολοκληρωμένης μεθοδολογίας για την ανίχνευση των τάσεων της αγοράς στον κλάδο των παιχνιδιών, θα προσδιορίσουμε όπως έχει προαναφερθεί ότι αναφερόμαστε σε συμβατικά και παραδοσιακά παιχνίδια που απευθύνονται σε παιδιά 0-14 ετών, χωρίς αυτό να είναι περιοριστικό.

Το target group σύμφωνα με το οποίο θα επιλεχθούν οι πηγές είναι τόσο τα παιδιά (κυρίως τα μεγαλύτερα) που εκφράζονται μέσα από τα κοινωνικά μέσα πλέον σε καθημερινή βάση, όσο και οι γονείς (όλων των παιδιών αυτή τη φορά και κυρίως των μικρότερων). Σχετικά με τα παιδιά, παρόλο που τυπικά η χρήση των πιο δημοφιλών κοινωνικών μέσων απαγορεύεται για χρήστες κάτω των 18 ετών, είναι εμφανές ότι αυτό δεν τηρείται. Περίπου το 60% των παιδιών έχουν χρησιμοποιήσει κάποιο κοινωνικό μέσο μέχρι την ηλικία των 10 ετών *(Πηγή: dailymail.co.uk, από έρευνα που πραγματοποιήθηκε σε δείγμα 1004 παιδιών ηλικίας από 8 έως 16 ετών και των γονέων τους).*

Σε μεγαλύτερες ηλικίες, τα μέσα κοινωνικής δικτύωσης παίζουν σημαντικό ρόλο στην καθημερινότητα των περισσότερων παιδιών και εφήβων. Πάνω από το 60% των ατόμων από 13 έως 17 ετών έχουν τουλάχιστον ένα προφίλ σε κάποιο κοινωνικό μέσο, ξοδεύοντας έως και πάνω από 2 ώρες σε αυτό σε καθημερική βάση *(Πηγή: American Academy of Child & Adolescent Psychiatry, έρευνα του 2017).*

Αντίστοιχα, οι ιστοσελίδες και τα ιστολόγια με θεματολογία που αφορούν τους γονείς, οι σελίδες δημόσιας συζήτησης, και τα κοινωνικά μέσα έρχονται αθροιστικά στην πρώτη και δεύτερη θέση σε ποσοστό 71% ως πηγές από τις οποίες επηρεάζονται οι γονείς της τρέχουσας γενιάς των millennials/generation Y *(Πηγή: Futurecast, "Millenials as new* parents," 2013). Πρόσφατη μελέτη της Crowdtap αποκάλυψε ότι το 97% των μαμάδων αυτής της γενιάς και το 93% των μπαμπάδων βρίσκουν τα κοινωνικά μέσα από λίγο έχως πολύ βοηθητικά στη γονική τους ιδιότητα και καθήκοντα.



Εικόνα 24: Τα δημοφιλέστερα κοινωνικά μέσα (έφηβοι και νέοι ενήλικες). Πηγή: statistica.com

	Facebook	Instagram	Pinterest	LinkedIn	Twitter
Total	68%	28%	26%	25%	21%
Men	67%	23%	15%	28%	21%
Women	69%	32%	38%	23%	21%
Ages 18-29	88%	59%	36%	34%	36%
30-49	79%	31%	32%	31%	22%
50-64	61%	13%	24%	21%	18%
65+	36%	5%	9%	11%	6%
High school or less	56%	19%	18%	9%	14%
Some college	77%	35%	31%	25%	24%
College graduate	77%	32%	33%	49%	28%
Less than $30,000	65%	29%	23%	16%	18%
$30,000-$49,999	68%	27%	27%	11%	16%
$50,000-$74,999	70%	30%	29%	30%	26%
$75,000+	76%	30%	34%	45%	30%
Urban	70%	34%	26%	29%	22%
Suburban	68%	24%	29%	26%	21%
Rural	65%	25%	20%	15%	19%

Εικόνα 25: Στατιστικά στοιχεία χρήσης των κοινωνικών μέσων (Πηγή:Pew Research center, Απρίλιος 2017)

Σε έρευνα που έγινε στις ΗΠΑ τον Φεβρουάριο του 2017 αναδείχθηκαν οι προτιμήσεις των εφήβων και νέων ενηλίκων (ηλικίας από 13 ετών) σχετικά με την παρουσία τους στα κοινωνικά μέσα (Εικόνα 23) ενώ κάποια γενικότερα στατιστικά επίσης από πρόσφατη έρευνα (Μάρτιος-Απρίλιος του 2017) φαίνονται στην Εικόνα 24. Γενικά, τα στατιστικά στοιχεία που μπορούμε να βρούμε σχετικά με τη χρήση των κοινωνικών μέσων είναι πολλά και συγκλίνουν ως προς τα συμπεράσματα που βγαίνουν από αυτά, καθώς η επισκεψιμότητα των ιστοσελίδων και η χρήση των εφαρμογών είναι στοιχεία εύκολα μετρήσιμα και με ανοιχτή πρόσβαση.

Συνοψίζοντας, τα δημοφιλέστερα κοινωνικά μέσα ανάμεσα στα οποία μπορούμε να επιλέξουμε ως πηγές είναι τα:

- Facebook
- Twitter
- LinkedIn
- Pinterest
- YouTube
- Instagram
- Tumblr
- Snapchat

ενώ οι διαδικτυακές πηγές με περιεχόμενο δημιουργημένο από χρήστες (user generated content) είναι τα ιστολόγια (Blogs), και οι ιστοσελίδες δημόσιας συζήτησης (Forums).

Λαμβάνοντας υπόψιν τη δημοτικότητα των μέσων στις ηλικιακές ομάδες, τη χρήση για την οποία προορίζεται το κάθε κοινωνικό μέσο αλλά και τους πρακτικούς περιορισμούς, επιλέξαμε να κατευθύνουμε τη συλλογή δεδομένων που αποτελεί το πρώτο στάδιο της μεθοδολογίας στο Facebook. Αυτό γιατί χρησιμοποιείται περισσότερο με διαφορά από τα υπόλοιπα μέσα από όλες τις ηλικίες, και περιλαμβάνει ομάδες, κοινότητες και σελίδες για κάθε θέμα, πλούσιες σε περιεχόμενο δημιουργημένο από τους χρήστες.

Το Twitter και το Linkedin απορρίφθηκαν ως πιο «σοβαρά» κοινωνικά μέσα, με το Twitter να έχει πιο πολύ ειδησεογραφικό χαρακτήρα και να ενδείκνυται περισσότερο για έκφραση απόψεων σε πιο αυστηρά ζητήματα και το Linkedin να προορίζεται μόνο για επαγγελματική χρήση που έχει λίγη ως μηδενική σχέση με το θέμα μας. Επίσης, το Twitter δείχνει μια πτωτική τάση ως προς τη δημοτικότητά του, έχοντας «μόλις» 317 εκ. χρήστες σε μηνιαία βάση, εκ των οποίων οι μισοί δεν δημοσιεύουν αναρτήσεις. Αντίθετα το Facebook διατηρεί την πρωτοκαθεδρία με 1.9 δις. ενεργούς μηνιαίους χρήστες, και με την ηλικιακή ομάδα 18-49 να περνάει περίπου 7 ώρες την εβδομάδα σε αυτό. Παρόλα αυτά, επειδή ο χειρισμός των αναρτήσεων των κοινωνικών δικτύων είναι σε μεγάλο βαθμό παρόμοιος, εύκολα μπορεί να προστεθεί στο στάδιο συλλογής δεδομένων και το API του Twitter εφόσον αυτό είναι επιθυμητό.

Λόγω πρακτικών περιορισμών, αποκλείσαμε τα δημοφιλή και ανερχόμενα μέσα τα οποία απαιτούν αλγορίθμους επεξεργασίας εικόνας και βίντεο στην εξόρυξη δεδομένων τους (π.χ. Youtube, Instagram). Το Instagram αποτελεί το γρηγορότερα αναπτυσσόμενο μέσο, με 600 εκ. μηνιαίους χρήστες, εκ των οποίων το 90% είναι κάτω των 35 ετών και το 53% «ακολουθεί» τουλάχιστον ένα λογαριασμό κάποιου «brand». Αυτό το καθιστά ένα εξαιρετικό κανάλι προώθησης για κάθε είδους προϊόντα, ειδικά για όσα απευθύνοται σε νεότερες ηλικίες (φοιτητές και νέους ενήλικες). Είναι όμως ένα μέσο του οποίου η παρακολούθηση με όρους Εξόρυξης Δεδομένων και Εξαγωγής Γνώσης είναι πρακτικά δύσκολη, διότι πέρα από τους αλγορίθμους επεξεργασίας εικόνας που αναφέρθηκαν, τα hashtags και οι λεζάντες που χρησιμοποιούνται από τους χρήστες είναι πολλές φορές από λίγο έως καθόλου αντιπροσωπευτικά του θέματος της εικόνας που έχει αναρτηθεί.

Τέλος, τα ιστολόγια και οι ιστοσελίδες δημόσιας συζήτησης, μπορεί να μη συγκρίνονται σε επισκεψιμότητα με τα μέσα κοινωνικής δικτύωσης, είναι όμως σημαντικές πηγές για τους γονείς

παιδιών μικρότερης ηλικίας, οι οποίοι καταφεύγουν σε αυτά ως πιο αξιόπιστες και εξειδικευμένες πηγές για να μοιραστούν τους προβληματισμούς τους και να πάρουν συμβουλές. Για το λόγο αυτό αποφασίσαμε να τα συμπεριλάβουμε σαν πηγές δεδομένων.

4.3.2. Επιλογή των σταδίων

Σε προηγούμενο κεφάλαιο, σαν **στάδια** ενός συστήματος ή μιας μεθοδολογίας Ανίχνευσης Τάσεων σε μέσα κοινωνικής δικτύωσης και διαδικτυακές πηγές έχουμε διακρίνει τα εξής:

1. **Συλλογή δεδομένων** (Data Collection)
2. **Επεξεργασία ή προ-επεξεργασία δεδομένων** (Data preprocessing)
3. **Ανίχνευση θέματος** (Topic Detection)
4. **Προσδιορισμός τάσης** (Trend Detection)
5. **Ανάλυση συναισθήματος** (Sentiment Analysis)

Τα στάδια αυτά όπως έχει προαναφερθεί είναι κοινά για τον εντοπισμό τάσεων, είτε πρόκειται για σημαντικά γεγονότα είτε για ανερχόμενα θέματα, και μπορούν να εξυπηρετήσουν διαφορετικούς σκοπούς στα πλαίσια κοινωνικής, ακαδημαϊκής και επιχειρηματικής έρευνας. Ορίζοντας σαν τάσεις της αγοράς για έναν κλάδο τα θέματα που είναι πιο δημοφιλή ανάμεσα στους καταναλωτές-χρήστες του διαδικτύου και είναι σχετικά με αυτόν, μπορούμε να προχωρήσουμε στη χρησιμοποίηση των σταδίων και των αντίστοιχων τεχνικών για την ανίχνευση τάσεων στον κλάδο που μελετάμε.

Σύμφωνα με τις πηγές δεδομένων που έχουμε επιλέξει, το πρώτο στάδιο θα πραγματοποιείται ξεχωριστά για το Facebook, και ξεχωριστά για το σύνολο των υπόλοιπων ιστοσελίδων που έχουν επιλεχθεί (blogs, forums). Η προσπέλαση του Facebook API γίνεται με τη βοήθεια ενός απλού αλγορίθμου που θα ανασύρει τις δημόσιες αναρτήσεις από όλο το γράφο οι οποίες θα πληρούν κάποια συγκεκριμένα χαρακτηριστικά. Η προσπέλαση των ιστοσελίδων γίνεται με τη βοήθεια ενός προγράμματος Ανίχνευσης Ιστού. Λόγω της διαφορετικής δομής και φύσης των δύο κατηγοριών δεδομένων, αντίστοιχα ξεχωριστά πρέπει να συνεχιστεί το στάδιο της επεξεργασίας, και ιδανικά στη συνέχεια να αναπαρασταθούν τα δεδομένα με έναν ενιαίο τρόπο προκειμένου να ενοποιηθούν για το στάδιο της ανίχνευσης θέματος και προσδιορισμού τάσης.

Και στις δύο περιπτώσεις, η αναζήτηση των τάσεων γίνεται για ένα κύριο, προκαθορισμένο θέμα και όχι γενικά στον κοινωνικό γράφο και τον ιστό. Το θέμα αυτό είναι ο ίδιος ο κλάδος και οι λέξεις-κλειδιά από τις οποίες αντιπροσωπεύεται. Επομένως, θα συλλεχθούν εκ των προτέρων οι αναρτήσεις και τα κείμενα που αφορούν ένα συγκεκριμένο θέμα (topic), προσδιορισμένο από λέξεις κλειδιά όπως “toys”, “kids”, “kids playing”, “playtime”, “funtime”, “best toys”, “best gift” κλπ.

Επομένως, στην περίπτωση των τάσεων της αγοράς για συγκεκριμένο κλάδο, όλα τα δεδομένα που συγκεντρώνονται είναι σχετικά μεταξύ τους διότι αφορούν τον κλάδο και έχουν συγκεντρωθεί με αυτή τη λογική. Αυτό που μένει, είναι να ανιχνευθούν τα πιο δημοφιλή υπο-θέματα, τα οποία αποτελούν τις τάσεις. Η αναζήτηση δημοφιλών λέξεων-κλειδιών στις επιλεγμένες αναρτήσεις αναμένεται να καταλήξει σε συχνή αναφορά των παιχνιδιών ή των κατηγοριών παιχνιδιών ή των ηρώων τα οποία είναι πιο δημοφιλή τη συγκεκριμένη χρονική στιγμή, δηλαδή αναφέρονται περισσότερες φορές και από άτομα με μεγαλύτερη επιρροή (αυτά μπορεί να είναι είτε χρήστες του Facebook είτε bloggers). Για παράδειγμα, συχνή αναφορά στις αναρτήσεις που εμπεριέχουν τις λέξεις

“toys”, “kids playing” κ.λ.π. θα μπορούσαν να έχουν λέξεις όπως “marvel”, “ironman”, “disney”, “superheroes”, “electronic”, “coloring” κ.ο.κ.

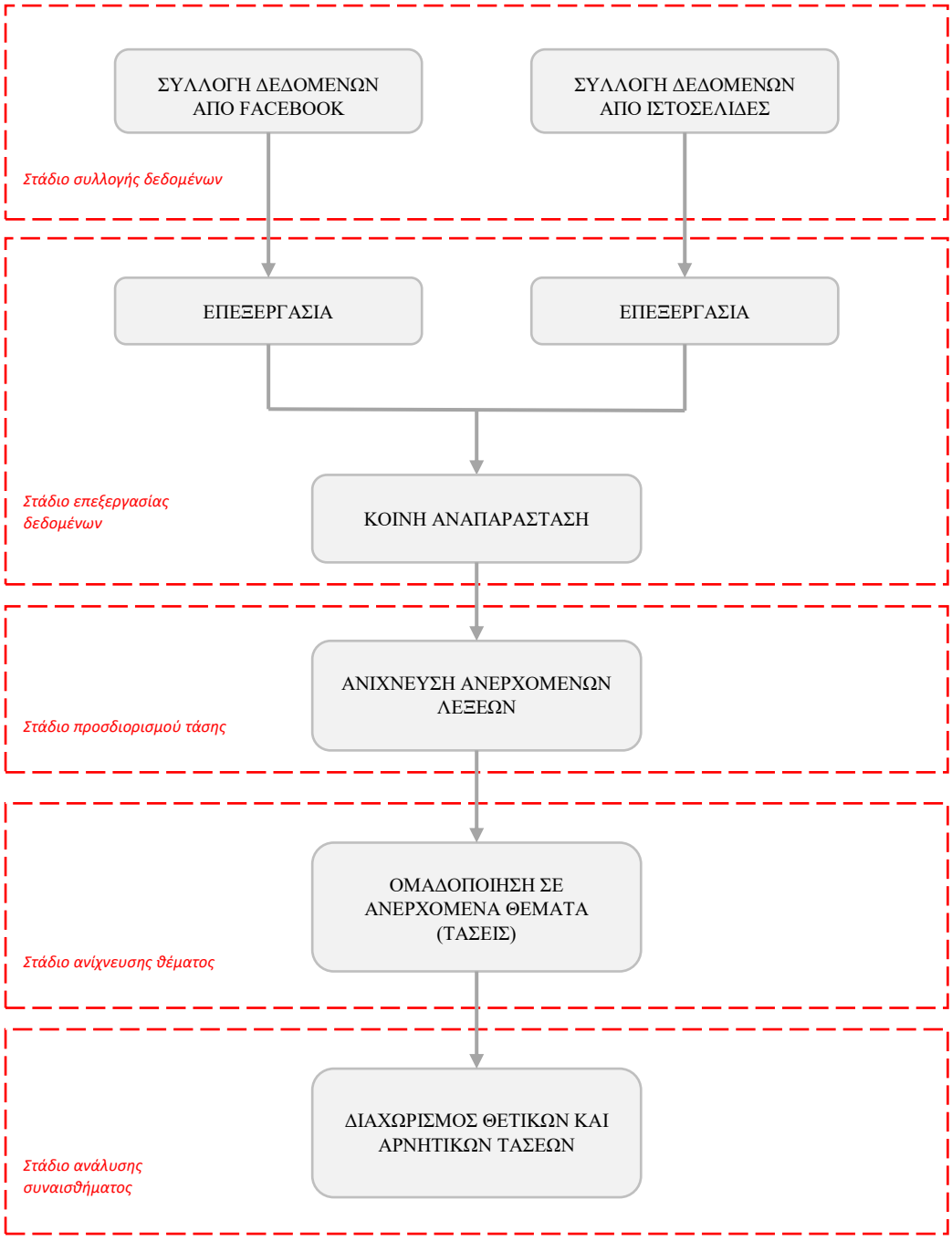
Άρα η λογική πορεία για την ανίχνευση τάσης στον κλάδο ξεκινά από τον προσδιορισμό των πηγών που αφορούν το target group παιδιών και γονέων, και συνεχίζει με την ξεχωριστή συλλογή των δεδομένων από τις πηγές, και συγκεκριμένα από τις αναρτήσεις που πληρούν κάποια θεματικά κριτήρια. Στη συνέχεια ακολουθεί η απαραίτητη επεξεργασία των αναρτήσεων και κειμένων που έχουν συλλεχθεί και η ενιαία αναπαράστασή τους. Τέλος, στα επεξεργασμένα δεδομένα ανιχνεύονται οι δημοφιλείς όροι ή λέξεις-κλειδιά και ομαδοποιούνται σε ευρύτερες κατηγορίες-θέματα, οι οποίες αποτελούν και τις τάσεις του κλάδου έτσι όπως εκφράζονται από τους χρήστες του διαδικτύου. Πρόκειται δηλαδή για μια μεθοδολογία ανίχνευσης ανερχόμενων θεμάτων, όπου το στάδιο ανίχνευσης θέματος συνυπάρχει με αυτό του προσδιορισμού τάσης.

Σχετικά με το στάδιο ανάλυσης συναισθήματος, από την έρευνα φάνηκε ότι είναι το σπανιότερο στις μεθόδους ανίχνευσης τάσης. Αυτό μπορεί να αιτιολογηθεί επειδή όταν ένα θέμα είναι ανερχόμενο, μπορεί να έχει τόσο θετική όσο και αρνητική χροιά. Σε κάποιες περιπτώσεις μάλιστα, ειδικά όταν πρόκειται για την ανίχνευση σημαντικών γεγονότων, είναι αναμενόμενο την προσοχή του κοινού να συγκεντρώνουν σε μεγάλο βαθμό αρνητικά γεγονότα.

Το συγκεκριμένο στάδιο θα εξυπηρετούσε σίγουρα ένα σύστημα επιχειρηματικής ευφυΐας για την παρακολούθηση των αντιδράσεων των καταναλωτών για κάποιο προϊόν ή τη δημοτικότητα της ίδιας της επιχείρησης και την εικόνα της στα μέσα κοινωνικής δικτύωσης και το διαδίκτυο. Μπορεί όμως αν ενσωματωθεί σε ένα σύστημα εντοπισμού τάσεων, να υποδείξει μοτίβα που έχουν τραβήξει την προσοχή των καταναλωτών αλλά που όμως έχουν καθαρά αρνητικό συναισθηματικό πρόσημο. Έτσι, μπορεί για παράδειγμα να αποφευχθεί η αναγνώριση μιας καινούριας ταινίας η οποία όμως αποδοκιμάστηκε από γονείς και παιδιά ως ενδεχόμενη θετική τάση για την κατασκευή παιχνιδιών με αντίστοιχη θεματολογία κ.ο.κ.

Σε αυτά τα πλαίσια, αποφασίσαμε να συμπεριλάβουμε το στάδιο ανάλυσης συναισθήματος μετά τον εντοπισμό των ανερχόμενων θεμάτων, ώστε να διερευνάται αν η τάση που εντοπίζεται είναι εν τέλει θετική ή αρνητική. Αυτό μπορεί να γίνει με μια τελική αναζήτηση στα δεδομένα που έχουν συλλεχθεί έτσι ώστε να προσδιοριστεί για κάθε θέμα πόσες φορές αναφέρεται με αρνητικά και πόσες με θετικά συμφραζόμενα.

Μια σχηματική απεικόνιση των σταδίων όπως έχουν αναλυθεί φαίνεται στο ακόλουθο σχήμα:



Εικόνα 26: Διάγραμμα ροής της προτεινόμενης μεθοδολογίας

4.3.3. Υλοποίηση των σταδίων

4.3.3.1. Συλλογή δεδομένων

Η διεπαφή προγραμματισμού του Facebook (Facebook Graph API) παρέχει πρόσβαση για τους εγγεγραμένους προγραμματιστές σε όλο τον γράφο του κοινωνικού δικτύου (χρήστες, αναρτήσεις, σελίδες κλπ). Για τους σκοπούς της παρούσας μεθοδολογίας ενδιαφερόμαστε κυρίως για τις αναρτήσεις. Κάθε ανάρτηση περιέχει τις εξής πληροφορίες [17]:

- Λεπτομέρειες για το περιεχόμενο της αναρτησης (μήνυμα, όνομα, περιγραφή)
- Τον χρήστη του Facebook ο οποίος δημοσίευσε το μήνυμα
- Το είδος του μηνύματος όπως προσδιορίζεται από το ίδιο το Facebook (κατάσταση/status, φωτογραφία, σύνδεσμος)
- Την ημερομηνία δημοσίευσης
- Την εφαρμογή που χρησιμοποιήθηκε για τη δημισίευση της ανάρτησης (π.χ. Facebook for Android)

Για την υλοποήση του σταδίου θα χρησιμοποιηθεί ένας απλός αλγόριθμος αναζήτησης ο οποίος θα ανασύρει όλες τις δημόσιες αναρτήσεις για κάθε λέξη του συνόλου λέξεων-κλειδιών που προσδιορίζει τον κλάδο. Δημοφιλείς γλώσσες προγραμματισμού για την υλοποίηση τέτοιων αλγορίθμων είναι η R, η Python και η Java. Το μέγιστο των αναρτήσεων που το API επιτρέπει να επιστρέφουν σε κάθε αναζήτης είναι 500. Σε περίπτωση που το αποτέλεσμα δεν είναι εξαντλητικό, η απάντηση που επιστρέφει το API σε μορφή JSON περιέχει σύνδεσμο για τη συνέχεια της αναζήτησης.

Ο παρακάτω αλγόριθμος (Εικόνα 26, [17]) αποτελεί παράδειγμα ενός τέτοιου απλού αλγορίθμου, σε ψευδοκώδικα. Ο συγκεκριμένος ανασύρει όλες τις αναρτήσεις, κάνοντας χρήση της επανάληψης όχι για κάποιες λέξεις-κλειδιά αλλά για κάθε χαρακτήρα του συστήματος ASCII.

```
/*Algorithm 1: Collection of public posts from Facebook*/

until = getLastCollectionTime();
foreach asciiChar in asciiList do
nextURL = collectPosts (asciiChar, until);
until (nextURL != null)
```

Εικόνα 27: Αλγόριθμος σε ψευδοκώδικα που ανασύρει όλες τις δημόσιες αναρτήσεις από το Facebook API

Στη συγκεκριμένη περίπτωση, το διάστημα αναμονής για την επανεκτέλεση του αλγορίθμου και την ανάσυρση των αναρτήσεων ορίστηκε στα 10 λεπτά. Το χρονικό αυτό διάστημα επιλέχθηκε ώστε να αφήνει ένα χρονικό κενό για την ανανέωση των δεδομένων αλλά ταυτόχρονα να είναι σχεδόν σε πραγματικό χρόνο.

Δεδομένου ότι στην παρούσα εργασία θα γίνει μια συγκεκριμένη αναζήτηση με λέξεις κλειδιά από ένα προκαθορισμένο σύνολο, τα αποτελέσματα του αλγορίθμου θα είναι λιγότερα και θα ανανεώνονται πιο αραιά. Επομένως το διάστημα θα πρέπει να οριστεί σχετικά μεγαλύτερο, ενδεχομένως της τάξης λίγων ωρών.

Σχετικά με την συλλογή δεδομένων από τις υπόλοιπες διαδικτυακές πηγές, θα χρησιμοποιηθεί κάποιος αλγόριθμος ανίχνευσης ιστού (crawler). Ο αλγόριθμος αυτός θα κάνει προσπέλαση ενός

συνόλου επιλεγμένων πηγών (ιστολόγια και σελίδες δημόσιας συζήτησης) και θα εξάγει από αυτές τα κείμενα που περιλαμβάνουν. Ένα παράδειγμα τέτοιου αλγορίθμου σε ψευδοκώδικα (Εικόνα 27, [23]) είναι το παρακάτω, το οποίο συλλέγει όλα τα κείμενα από ένα σύνολο πηγών και τα τοποθετεί σε μια λίστα.

```
/*Algorithm 2: Document Retrieval*/
/*Input: f_list, list of feeds Output: d_list, list of documents*/

d_list = null;
foreach f in f_list do
    d =  retrieveDocument(f);
    d_list.add(d);
return d_list;
```

Εικόνα 28: Αλγόριθμος σε ψευδοκώδικα που ανασύρει όλα τα κείμενα από ένα σύνολο ιστοσελίδων

4.3.3.2. Επεξεργασία δεδομένων

Δεδομένα που προέρχονται από το Facebook

Στα δεδομένα που προέρχονται από το Facebook επιλέξαμε να εφαρμόσουμε κάποιες από τις συνηθέστερες τεχνικές επεξεργασίας φυσικής γλώσσας οι οποίες στοχεύουν στην απαλλαγή των αναρτήσεων των κοινωνικών μέσων από περιττά στοιχεία. Στόχος είναι να μείνει όσο το δυνατόν η ουσία της ανάρτησης και οι αλγόριθμοι εντοπισμού των δημοφιλών λέξεων να αναζητούν λέξεις κλειδιά μέσα από ένα σύνολο στοιχείων της γλώσσας που είναι διαφορετικά μεταξύ τους και σημασιολογικά σημαντικά. Οι τεχνικές αυτές είναι:

- Αφαίρεση κοινών λέξεων (stop-words)
- Αφαίρεση εξωτερικών συνδέσμων (hyperlinks)
- Αφαίρεση αναφορών σε χρήστες (user mentions)
- Αναγωγή λέξεων στη σημασιολογική τους ρίζα (stemming)

Η επεξεργασία αυτή είναι ιδιαίτερα διαδεδομένη και μπορεί να πραγματοποιηθεί με τη βοήθεια εργαλείων ή και βιβλιοθηκών ανοιχτού κώδικα όπως είναι το Stanford's Core NLP Suite υλοποιημένο σε Java, το NLTK σε Python και Apache Lucene και OpenNLP.

Από την άλλη, επιλέξαμε να μην εξαιρέσουμε αναρτήσεις που περιέχουν λιγότερες λέξεις από έναν προκαθορισμένο αριθμό. Η επεξεργασία αυτή συναντάται συχνά στις μεθόδους που αφορούν τα κοινωνικά μέσα σύντομων αναρτήσεων (microblogs), όμως η φύση του μέσου επιτρέπει την ύπαρξη σύντομων αναρτήσεων που έχουν όμως νόημα. Για παράδειγμα, αναρτήσεις που συνοδεύουν κάποια εικόνα μπορεί να αποτελούνται απλά από λίγα hashtags ή λέξεις-κλειδιά που όμως να αναφέρουν περιεκτικά και το αντικείμενο της εικόνας αλλά και το συναίσθημα του χρήστη ως προς αυτό.

Επίσης, δεν χρειάζεται να εξαιρέσουμε αναρτήσεις που είναι σε άλλη γλώσσα διότι η αναζήτηση είναι ήδη περιορισμένη σε αναρτήσεις που περιέχουν λέξεις-κλειδιά οι οποίες είναι στην επιθυμητή γλώσσα, αλλά και γιατί η αγορά των παιχνιδιών όπως και οι κλάδοι που σχετίζονται με αυτή (ταινίες, σειρές, κινούμενα σχέδια, εικονογραφημένα περιοδικά και βιβλία) είναι διεθνής. Ακόμη και μια

ανάρτηση σε ξένη γλώσσα μπορεί να περιέχει το όνομα κάποιου ήρωα, μία από τις λέξεις-κλειδιά στα αγγλικά και ακόμα και κάποια λέξη έκφρασης συναισθήματος επίσης στα αγγλικά.

Τέλος, δεν θα εξαιρέσουμε σύμβολα και εικονίδια (emoticons) γιατί θα χρησιμεύσουν στο στάδιο της ανάλυσης συναισθήματος.

Δεδομένα που προέρχονται από ιστοσελίδες

Η επεξεργασία των δεδομένων που προέρχονται από τα ιστολόγια και τις ιστοσελίδες δημόσιας συζήτησης είναι αντίστοιχη, αφού στα κείμενα εφαρμόζονται οι ίδιες τεχνικές επεξεργασίας φυσικής γλώσσας ανεξαρτήτως μήκους ανάρτησης. Επομένως θα επαναλάβουμε τα τρία από τα τέσσερα στάδια στάδια που εφαρμόσαμε για την επεξεργασία των μικρών αναρτήσεων και θα προσθέσουμε τρία επιπλέον για να μειωθεί η πολυπλοκότητα και ο θόρυβος που έχουν τα μεγαλύτερα κείμενα. Συνολικά, τα στάδια επεξεργασίας για τα κείμενα των ιστοσελίδων είναι:

- Αφαίρεση των HTML στοιχείων που υπάρχουν στις ιστοσελίδες (εικόνες, μορφοποίηση)
- Αφαίρεση των εσωτερικών συνδέσμων περιήγησης (navigation links)
- Τμηματοποίηση των HTML κόμβων
- Αφαίρεση κοινών λέξεων (stop-words)
- Αφαίρεση εξωτερικών συνδέσμων (hyperlinks)
- Αναγωγή λέξεων στη σημασιολογική τους ρίζα (stemming)

Την παράλληλη συλλογή και επεξεργασία των δεδομένων που έχουν συλλεχθεί από το Facebook και το Twitter ακολουθεί η ενσωμάτωσή τους (data integration) σε μια κοινή βάση δεδομένων στην οποία μπορούν να περαστούν ως αρχείο (batch) όπως έχει γίνει και στην υλοποίηση του Cloud4Trends [26] αντίστοιχα για δεδομένα από το Twitter και μια συλλογή ιστοσελίδων.

4.3.3.3. Προσδιορισμός τάσης

Όπως έχουμε αναφέρει κατά την επιλογή των σταδίων, για τον προσδιορισμό τάσης θα ακολουθηθεί η μέθοδος ανεύρεσης των δημοφιλών όρων, ή αλλιώς των όρων που αναφέρονται συχνότερα στις επιλεγμένες αναρτήσεις. Για τους δημοφιλείς όρους, η κοινή προσέγγιση που μπορεί να ακολουθηθεί είναι η απόδοση βάρους στις λέξεις των αναρτήσεων, και η κατάταξή τους με σειρά δημοτικότητας.

Το αποτέλεσμα που προκύπτει από τη συλλογή και την επεξεργασία των δεδομένων από τα δύο είδη πηγών, έχει ενσωματωθεί από το προηγούμενο στάδιο σε μια κοινή βάση δεδομένων με ενιαία μορφή. Κατά την εισαγωγή, οι εγγραφές θα περιέχουν ενδεικτικά την εξής πληροφορία στις αντίστοιχες στήλες:

- Αύξοντα αριθμό
- Το επεξεργασμένο κείμενο της ανάρτησης ή της δημοσίευσης
- Ένα δείκτη που δηλώνει αν πρόκειται για ανάρτηση ή δημοσίευση
- Στην περίπτωση που πρόκειται για ανάρτηση, τον όρο αναζήτησης (λέξη-κλειδί) που έχει χρησιμοποιηθεί κατά την ανάσυρσή της

Οι τελευταίες δύο πληροφορίες θα χρησιμεύσουν ώστε να μπορεί να εφαρμοστεί η κατάλληλη παραλλαγή της TF-IDF προσέγγισης (κεφάλαιο 3.3.4.4) για τη συνένωση των μικρών αναρτήσεων σε

ψευδο-κείμενα, κάνοντας χρήση του κοινού όρου σύμφωνα με τον οποίο έχουν ανασυρθεί. Με αυτό τον τρόπο η TF-IDF μπορεί να εφαρμοστεί σε ένα βήμα και για τα δύο είδη πηγών και να αποδοθεί βάρος στις λέξεις χωρίς να υπάρχει το εμπόδιο των αραιών συνδέσμων περιεχομένου (sparse context links) για τις καταχωρήσεις που προέρχονται από το Facebook.

Το αποτέλεσμα του σταδίου Προσδιορισμού Τάσης περιλαμβάνει τη λίστα των Ανερχόμενων Όρων (trending terms) οι οποίοι στη συνέχεια θα ενοποιηθούν σε Ανερχόμενα Θέματα (trending topics).

4.3.3.4. Ανίχνευση θέματος

Για τους σκοπούς της προτεινόμενης μεθοδολογίας, η Ανίχνευση Θέματος γίνεται μαζί με τον Προσδιορισμό Τάσης ο οποίος φαινομενικά να προηγείται. Οι τάσεις έχουν ουσιαστικά προσδιοριστεί με τον εντοπισμό των ανερχόμενων όρων, επομένως τα θέματα που ανιχνεύονται στη συνέχεια είναι απευθείας τα δημοφιλή/ανερχόμενα θέματα. Το στάδιο αποτελείται από την ομαδοποίηση των αποτελεσμάτων του προηγούμενου σταδίου, και για το λόγο αυτό θα χρησιμοποιηθεί τεχνική Ομαδοποίησης και όχι Μοντέλου Θέματος.

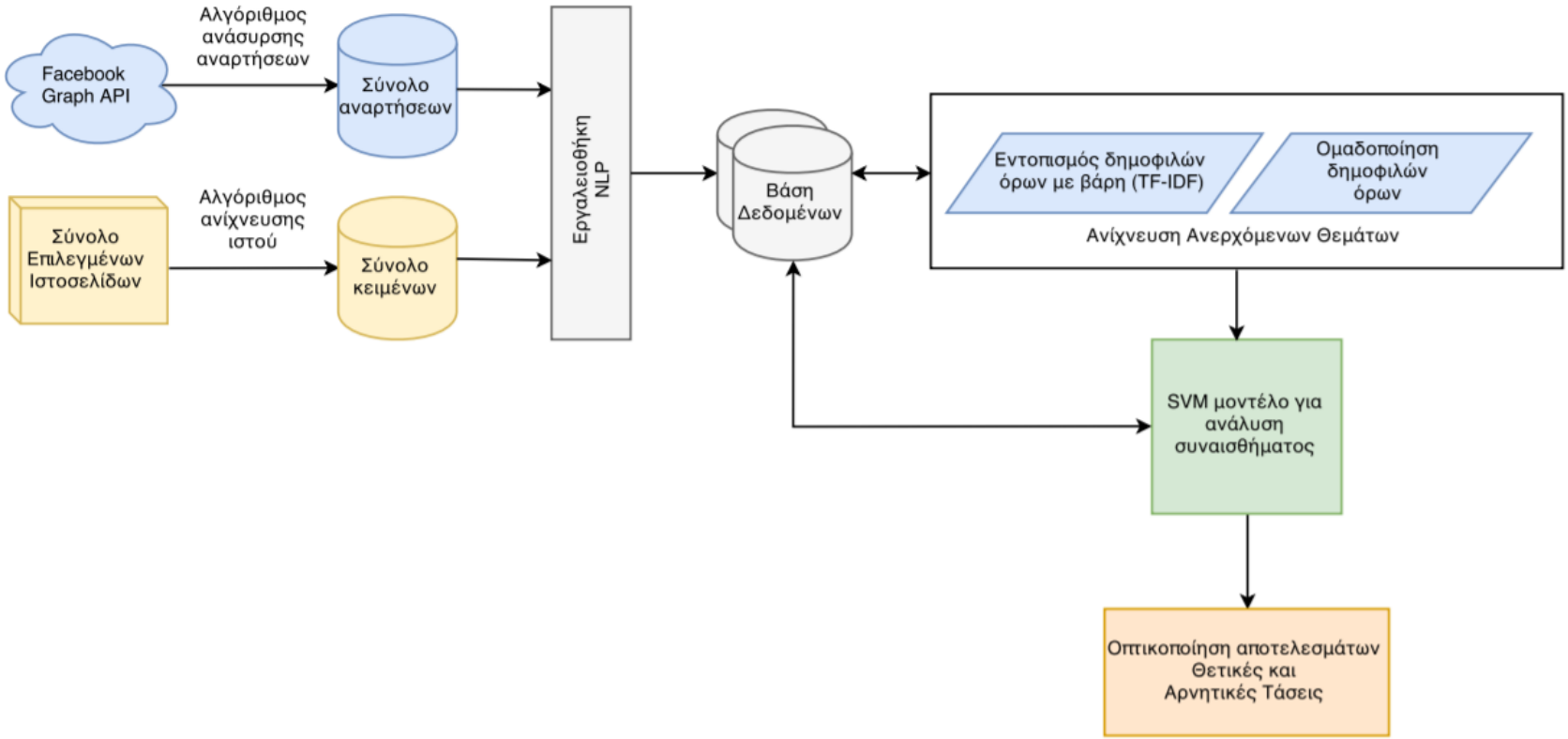
Επειδή η μεθοδολογία περιέχει και δεδομένα που προέρχονται από πηγές εκτός μέσων κοινωνικής δικτύωσης (μεικτά δεδομένα), οι μέθοδοι βασισμένες σε χαρακτηριστικά (feature-based approaches) απορρίπτονται αφού επικεντρώνονται σε χαρακτηριστικά στοιχεία των αναρτήσεων (π.χ. hashtags, χρονική σήμανση, χρήστες) και κάνουν μια ομαδοποίηση των χαρακτηριστικών αυτών και όχι των αναρτήσεων. Επομένως θα χρησιμοποιηθεί μέθοδος ομαδοποίησςη **βασισμένη σε λέξεις-κλειδιά (keyword-based approaches),** όπου η ανίχνευση θέματος πραγματοποιείται σύμφωνα με τη συχνότητα κοινής εμφάνισής τους και αφού έχει προηγηθεί υπολογισμός της σημαντικότητας/βάρους των λέξεων-κλειδιών ή δημοφιλών όρων ([30], [31], [17], [38], [34]).

Επίσης, Ομαδοποίηση μπορεί να γίνει προαιρετικά και σε επίπεδο θέματος όπου δημοφιλή υπο-θέματα που έχουν ανιχνευθεί ομαδοποιούνται σε ένα μεγαλύτερο ([40], [34], [41]). Αυτό μπορεί να γίνει, για παράδειγμα, αν το σύνολο των θεμάτων που έχουν ανιχνευθεί ξεπερνά ένα ορισμένο κατώφλι ανά αριθμό αναρτήσεων.

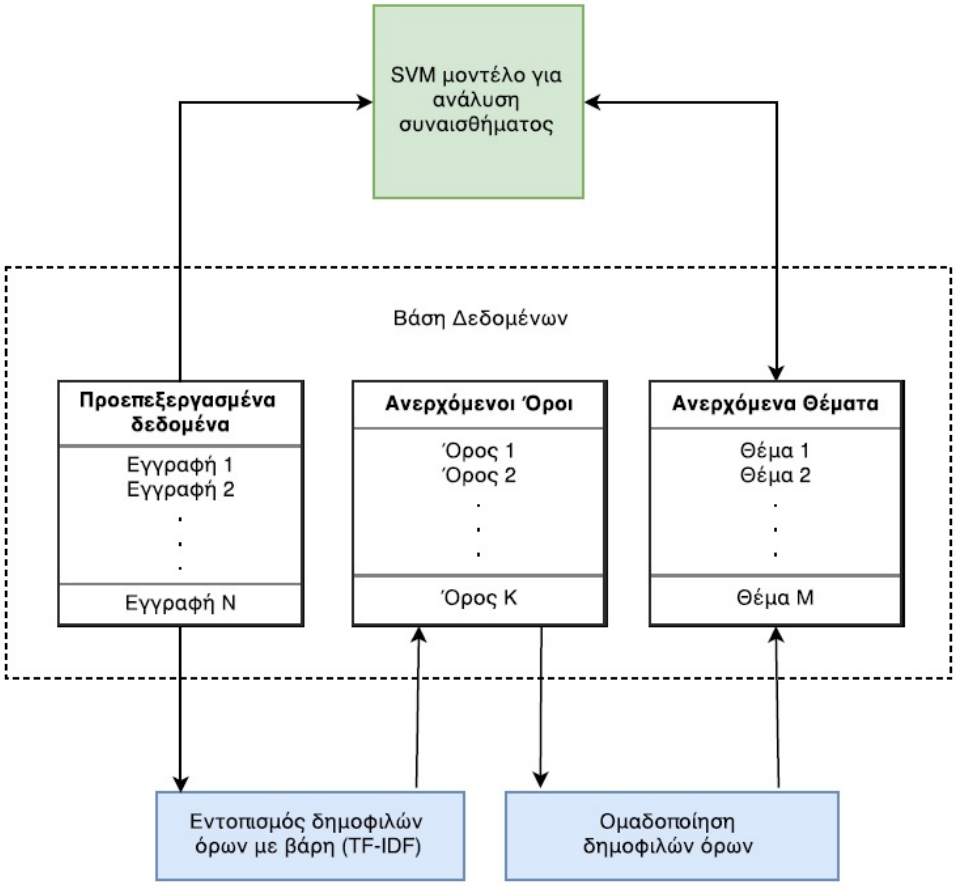
4.3.3.5. Ανάλυση συναισθήματος

Για το τελευταίο στάδιο της μεθοδολογίας, το οποίο θα αποδώσει συναισθηματικό προσανατολισμό στα ανερχόμενα θέματα που έχουν εντοπιστεί, επιλέγουμε την υλοποίηση με μοντέλο Μηχανής Υποστήριξης Διανυσμάτων (SVM από το Support Vector Machine). Πρόκειται για μοντέλο επιβλεπόμενης μάθησης, όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο, με αλγορίθμους μάθησης που αναλύουν δεδομένα που χρησιμοποιούνται για ταξινόμηση και ανάλυση παλινδρόμησης.

Στην περίπτωση χρήσης SVM για ανάλυση συναισθήματος, ο χώρος που αποτελείται από σημεία – αναρτήσεις, χωρίζεται σε δύο τμήματα, ένα για κάθε συναίσθημα (θετικό, αρνητικό). Το SVM προσπαθεί να κατατάξει τις αναρτήσεις σε κάθε πλευρά (σε κάθε συναίσθημα). Δημοφιλείς βιβλιοθήκες για την υλοποιηση SVM αλγορίθμων υπάρχουν σε πολλές σύγχρονες γλώσσες προγραμματισμού όπως η Python, η Ruby, η Java και η R. Τέλος, κατά την υλοποίηση θα πρέπει να προβλεφθεί η συλλογή ενός συνόλου δεδομένων για την εκπαίδευση του αλγορίθμου (training dataset).



Εικόνα 29: Σχηματική αναπαράσταση της προτεινόμενης μεθοδολογίας



Εικόνα 30: Σχηματική αναπαράσταση των περιεχομένων της βάσης

5. Συμπεράσματα και προτάσεις

Στο πλαίσιο της παρούσας διπλωματικής εργασίας μελετήθηκε η δυνατότητα ανίχνευσης των τάσεων της αγοράς στα κοινωνικά μέσα και τον ιστό του διαδικτύου. Η έρευνα αγοράς και προτιμήσεων καταναλωτή εξετάστηκε αρχικά σε θεωρητικό επίπεδο, από τη σκοπιά του κλάδου του μάρκετινγκ, και δόθηκε μια γενική εικόνα των παραδοσιακών μεθόδων αλλά και των σύγχρονων κατευθύνσεων στον τομέα της έρευνας αγοράς. Στη συνέχεια, μελετήθηκε η ανίχνευση των τάσεων της αγοράς από διαδικτυακές πηγές από τεχνική σκοπιά και έγινε αναφορά στα βασικά σημεία του γενικότερου τεχνικού υπόβαθρου που απαιτείται για την κατανόηση του ερευνητικού αντικειμένου και τη συμβολή σε αυτό (Επιχειρηματική Ευφυΐα, Εξόρυξη Δεδομένων, Μεγάλα Δεδομένα).

Πραγματοποιήθηκε εκτεταμένη βιβλιογραφική επισκόπηση των μεθόδων που αφορούν το θέμα μας, δηλαδή την κατασκευή συστημάτων Επιχειρηματικής Ευφυΐας βασισμένων σε μεθόδους Εξόρυξης Δεδομένων, ειδικά όμως για την ανίχνευση των τάσεων της αγοράς. Για τους σκοπούς της έρευνας, μελετήθηκαν πάνω από 50 μεθοδολογίες, 22 εκ των οποίων τα χαρακτηριστικά αποτυπώθηκαν στον σχετικό πίνακα στο παράρτημα της εργασίας. Το δείγμα αυτό θεωρείται αντιπροσωπευτικό του συνόλου.

Για την αποτύπωση των μεθοδολογιών στον πίνακα, πραγματοποιήθηκε η κατηγοριοποίηση και η ανάλυση των σταδίων της διαδικασίας Ανίχνευσης Τάσης όπως έχουν προσδιοριστεί στο κεφάλαιο 3. Ο πίνακας περιέχει 22 μεθοδολογίες Ανίχνευσης Τάσης οι οποίες είτε έχουν υλοποιηθεί ως μέρη ολοκληρωμένου συστήματος είτε αποτελούν προτάσεις για υλοποίηση. Έχουν επιλεχθεί από το σύνολο της βιβλιογραφίας που μελετήθηκε ως ενδεικτικές στην κατηγορία τους, και με κριτήριο να έχει πραγματοποιηθεί πειραματικός έλεγχος σε τουλάχιστον ένα πραγματικό σύνολο δεδομένων.

Οι δύο πρώτες στήλες του πίνακα περιέχουν το όνομα και την χρονολογία δημοσίευσης. Στη συνέχεια ακολουθούν κάποια βασικά χαρακτηριστικά της μεθοδολογίας που περιγράφεται: Αν έχει εφαρμοστεί (Έτοιμο εργαλείο), για ποιες πηγές ή/και κοινωνικά μέσα έχει δημιουργηθεί, σε ποια κατηγορία Ανίχνευσης Τάσης αναφέρεται (Εντοπισμό ανερχόμενων θεμάτων, Εντοπισμό σημαντικών γεγονότων, Ανίχνευση τάσεων για προκαθορισμένα θέματα), αν κάνει έγκαιρη ανίχνευση τάσης και αν μπορεί να εφαρμοστεί σε πραγματικό χρόνο. Στις επόμενες δύο στήλες περιέχονται η είσοδος και η έξοδος των πειραμάτων ή του εργαλείου σε περίπτωση που πρόκειται για υλοποιημένη μέθοδο.

Τέλος, ακολουθούν τα στάδια που υπάρχουν στην κάθε μεθοδολογία, καθώς και οι τεχνικές που χρησιμοποιούνται στο κάθε ένα από αυτά. Τα βήματα της κάθε μεθοδολογίας σε πιο αναλυτική μορφή εμπεριέχονται σε ξεχωριστό πίνακα. Οι δύο πίνακες είναι τοποθετημένοι σε Παράρτημα στο τέλος της εργασίας.

Σαν αποτέλεσμα της ερευνητικής διαδικασίας, και αφού πραγματοποιήθηκε έρευνα σχετικά με τις ανάγκες του κλάδου της βιομηχανίας κατασκευής παιχνιδιών, προτάθηκε μια μεθοδολογία για την ανίχνευση των τάσεων της αγοράς για τον συγκεκριμένο κλάδο. Επιλέχθηκε ο κατάλληλος συνδυασμός σύγχρονων τεχνικών για τον εντοπισμό των τάσεων της αγοράς στο συγκεκριμένο κλάδο μέσα από τα κοινωνικά δίκτυα και άλλες διαδικτυακές πηγές. Από το συνδυασμό των τεχνικών και τη ροή της μεθόδου όπως αποτυπώθηκε στο Κεφάλαιο 4 της εργασίας θα μπορεί να σχεδιαστεί ένα ολοκληρωμένο σύστημα ανίχνευσης τάσεων, το οποίο θα χρησιμεύσει σαν ένα εργαλείο επιχειρηματικής ευφυΐας στα χέρια μια εταιρίας αυτού του τομέα για την αναγνώριση πιθανών «κενών» ή ευκαιριών της εξεταζόμενης αγοράς, προς τα οποία μπορεί να προσανατολιστεί.

Συνοψίζοντας, τα αποτελέσματα της έρευνας συγκεντρώνονται σε δύο άξονες. Πρώτον, γίνεται μια κατηγοριοποίηση των μεθοδολογιών ανίχνευσης τάσεων και των βημάτων που αυτές ακολουθούν, καθώς και των τεχνικών που χρησιμοποιούνται. Δεύτερον, στο τελευταίο μέρος της εργασίας εξετάζεται κατά πόσο είναι χρήσιμη η εξόρυξη δεδομένων από τα κοινωνικά δίκτυα και το διαδίκτυο για τον κλάδο των παιχνιδιών, και προτείνεται ένα ολοκληρωμένο σύστημα ανίχνευσης των τάσεων της αγοράς από τις πηγές αυτές για το συγκεκριμένο κλάδο.

Σύνοψη της διαδικασίας που ακολουθήθηκε:

- ✓ Μελέτη της τεχνολογικής προόδου στον χώρο της ανίχνευσης τάσεων
- ✓ Εκτεταμένη βιβλιογραφική αναζήτηση
- ✓ Συγκέντρωση των πιο χαρακτηριστικών δημοσιεύσεων σύμφωνα με το θέμα
- ✓ Κατηγοριοποίησή και εντοπισμός των βασικών σταδίων
- ✓ Διερεύνηση των επιμέρους τεχνικών που χρησιμοποιούνται σε κάθε στάδιο
- ✓ Ανάλυση και σύγκριση των χαρακτηριστικών των υφιστάμενων προσεγγίσεων
- ✓ Συνοπτική μελέτη του κλάδου της βιομηχανίας παιχνιδιών
- ✓ Πρόταση μεθοδολογίας ανίχνευσης τάσεων τον συγκεκριμένο κλάδο

Σύνοψη Συμπερασμάτων:

- ✓ Οι **μεθοδολογίες Ανίχνευσης Τάσεων** στο διαδίκτυο χρησιμοποιούν παρόμοιες τεχνικές ανεξάρτητα από την εξειδίκευσή τους και τον ορισμό της τάσης που δίνει ο ερευνητής
- ✓ Με βάση την έρευνα που έγινε καταλήξαμε σε **5 βασικά στάδια** τα οποία είναι κοινά σε μια μεθοδολογία Ανίχνευσης Τάσης και διαφοροποιούνται ανάλογα με τη στόχευση
- ✓ Από συνδυασμό των σταδίων και των τεχνικών σχεδιάστηκε ένα ολοκληρωμένο **σύστημα Ανίχνευσης Τάσεων της Αγοράς** για τον κλάδο της Βιομηχανίας Παιχνιδιών, το οποίο μπορεί να χρησιμεύσει σαν εργαλείο επιχειρηματικής ευφυΐας στα χέρια μια εταιρίας του κλάδου

Σύνοψη προοπτικών για μελλοντική έρευνα:

- ✓ Υλοποίηση του συστήματος **Ανίχνευσης Τάσεων της Αγοράς** που προτάθηκε
- ✓ Επέκταση του σταδίου **Προσδιορισμού Τάσης** ώστε να γίνεται καλύτερη εκμετάλλευση των χαρακτηριστικών των αναρτήσεων και του γράφου του κοινωνικού δικτύου (π.χ. δημοτικότητα ανάρτησης, προφίλ χρήστη)
- ✓ Ενσωμάτωση περισσότερων **πηγών δεδομένων** που είναι ανερχόμενες και έχουν περιεχόμενο που αφορά τον κλάδο (π.χ. Instagram, Snapchat) με χρήση αλγορίθμων επεξεργασίας εικόνας και βίντεο

6. Παράρτημα

Πίνακας 5: Βιβλιογραφική επισκόπηση μεθοδολογιών ανίχνευσης τάσεων – Χαρακτηριστικά/Εφαρμογή

	Δημοσίευση		Χαρακτηριστικά					Πείραμα/εφαρμογή	
Α/Α	**Τίτλος**	**Χρονολογία**	**Έτοιμο εργαλείο**	**Πηγές/Πεδίο εφαρμογής**	**Αντικείμενο εφαρμογής**	**Ανίχνευση σε αρχικό στάδιο**	**Σε πραγματικό χρόνο**	**Σύνολο Δεδομένων**	**Έξοδος**
1	BlogPulse: Automated Trend Discovery for Weblogs	2004	Ναι	Blogs	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Όχι	Δεν αναφέρεται πείραμα σε συγκεκριμένο σύνολο δεδομένων. Είσοδος του συστήματος είναι η λίστα των πηγών	Τα θέματα που ανιχνεύονται στο σώμα και η τάση τους στο χρόνο, με τη βοήθεια αυτοματοποιημένων αλγορίθμων
2	TwitterStand: News in Tweets	2009	Ναι	Twitter	Ανίχνευση σημαντικών γεγονότων	Όχι	Όχι	Σύνολο αναρτήσεων που έχουν συλλεχθεί από 4 διαφορετικές ροές (Seeders, GardenHose, Search, BirdDog)	Τα αποτελέσματα δίνονται μέσω της εφαρμογής του συστήματος και αποτελούνται από το αποτέλεσμα της ομαδοποίησης σε θέματα και τη γεωγραφική τους απεικόνοση σε χάρτη
3	Twittermonitor: trend detection over the twitter stream	2010	Ναι	Twitter	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	Το σύνολο δεδομένων όπου εφαρμόζεται η μεθοδολογία αποτελείται από ένα δείγμα αναρτήσεων στο Twitter της τάξης των 10 εκ. Ανά ημέρα.	Η έξοδος παρουσιάζεται στην διαδικτυακή εφαρμογή του εργαλείου, και είναι οι τάσεις που εντοπίζει σε πραγματικό χρόνο.

4	Monitoring trends on Facebook	2011	Όχι	Facebook	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	Σύνολο από 2.000.000 δημόσιες αναρτήσεις,σε διάστημα 4 ημερών, κατά το οποίο αναμενόταν να ανιχνευθούν τρία γνωστά σημαντικά γεγονότα. Πραγματοποιήθηκαν 10 πειράματα με 1000 αναρτήσεις το καθένα	Δύο εκ των τριών γεγονότων ανιχνεύθηκαν με επιτυχία ενώ το τρίτο λόγω της διαφορετικότητας των όρων οι οποίοι αναφέρονται σε αυτό ανιχνεύθηκε ελλιπώς
5	Social networking trends and dynamics detection via a cloud-based framework design (Cloud4Trends)	2012	Ναι	Blogs, Twitter	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	Τα δεδομένα συλλέγονται από το Twitter Streaming API και από το Google Blogger API σε πραγματικό χρόνο, μέσω parsers	Η έξοδος παρουσιάζεται στη διαδικτυακή εφαρμογή του εργαλείου, και είναι οι τάσεις που εντοπίζει
6	Microblog Topic Detection Based on LDA Model and Single-Pass Clustering	2012	Όχι	Microblogs	Ανίχνευση θεμάτων	Όχι	Ναι	Έλεγχος της μεθοδολογίας σε ένα σύνολο με 108.122 αναρτήσεις από το Sina-microblog τον Αύγουστο του 2011 με τη βοήθεια ενός προγράμματος ανίχνευσης ιστού	Τα θέματα στα οποία αναφέρονται οι αναρτήσεις της πειραματικής εισόδου είναι γνωστά από πριν, ώστε να ελεγχθεί η αποδοτικότητα της μεθόδου στην ανίχνευσή τους, σε σύγκριση και με το VSM μοντέλο
7	Open Domain Event Extraction from Twitter (TwiCal)	2012	Ναι	Twitter	Ανίχνευση σημαντικών γεγονότων	Όχι	Όχι	Σύνολο 100 εκ. αναρτήσεων του Twitter (11/2011), τα οποία συλλέχτηκαν από το Twitter Streaming API με χρήση αναζήτησης ενός συνόλου χρονικών λέξεων κλειδιών, όπως “today”, “tomorrow”, ονόματα ημερών, μηνών κοκ	Τα πιο σημαντικά γεγονότα για κάθε ευρύτερη κατηγορία με τη μορφή οντοτήτων , οι πιο χαρακτηριστικές φράσεις για το κάθε ένα. Παρουσιάζονται τα κορυφαία 5 ευρήματα

8	Deriving market intelligence from microblogs	2013	Όχι	Microblogs	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	Γίνονται δύο πειράματα με δεδομένα που έχουν συλλεχθεί από το Twitter με χρήση της αναζήτησης (querying) στο API του. Χρησιμοποιούνται αναζητήσεις σχετικά με τρείς εταιρείες και τρία προϊόντα.	Η έξοδος περιλαμβάνει τα δημοφιλή θέματα, το συναισθηματικό προσανατολισμό των αναρτήσεων που τα αφορούν και τα συμπεράσματα από τη συνολική εικόνα της αναζήτησης.
9	Discovering hot topics using Twitter streaming data social topic detection and geographic clustering	2013	Όχι	Κοινωνικά δίκτυα	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	18,720,902 γεωγραφικά προσημασμένες (geo-tagged) αναρτήσεις του Twitter στο διάστημα 23/3/2013-1/4/2013 στις ΗΠΑ. Ανίχνευση θέματος ανά γεωγραφική περιοχή	Προσδιορίστηκαν 9 ανερχόμενα θέματα με ημι-αυτόματο τρόπο και δημιουργήθηκαν γραφικές παραστάσεις με την τάση καθενός από αυτά. Τα αποτελέσματα ομαδοποιήθηκαν ανά πολιτεία και ανά περιοχή
10	Detecting trends on the Web: A multidisciplinary approach	2014	Όχι	Blogs, Twitter	Ανίχνευση σημαντικών γεγονότων	Όχι	Όχι	Ένα σύνολο από 200,890 αντικειμενικά κείμενα και 268,800 tweets	Ανάμεσα στα 117 θέματα 65 σημαντικά γεγονότα εντοπίστηκαν χειροκίνητα
11	Hashtag Graph based Topic Model for Tweet Mining	2014	Όχι	Twitter	Ανίχνευση θεμάτων	Όχι	Άγνωστο	Το έτοιμο dataset “TweetData”1, το οποίο περιέχει 16 εκ. αναρτήσεων από το Twitter που έχουν δειγματοληφθεί μεταξύ 23/01/2011 και 8/02/2011	Τα θέματα που ανακαλύπτει το μοντέλο και τα αντιπροσωπευτικά τους hashtags. Στο πείραμα ανιχνεύονται 60 θέματα
12	PoliTwi: Early detection of emerging political topics on twitter and the impact on	2014	Ναι	Twitter	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	Τα δεδομένα συλλέγονται από το Twitter Streaming API και αποθηκεύονται σε μια PostgreSQL βάση	Η έξοδος του συστήματος είναι σε μορφή παρουσίασης των κορυφαίων δημοφιλών θεμάτων

	concept-level sentiment analysis							πριν την προεπεξεργασία	μαζί με την ανάλυση συναισθήματος και γίνεται από διαφορετικά κανάλια (Twitter, ιστοσελίδα και εφαρμογή για κινητά)
13	A prerecognition model for hot topic discovery based on microblogging data	2014	Όχι	Microblogs	Ανίχνευση ανερχόμενων θεμάτων	Ναι	Ναι	Σύνολο 2,000,000 αναρτήσεων από το κοινωνικό δίκτυο Sina. Μετά το στάδιο της προεπεξεργασίας μένουν 675,439 αναρτήσεις στο διάστημα 2014/01/01 με 2014/04/30. Χρησιμοποιούνται κάποια δεδομένα σαν σύνολο μάθησης και τα υπόλοιπα για έλεγχο	Η μεθοδολογία επιστρέφει ταξινομημένα τα πιο δημοφιλή θέματα που έχουν ανιχνευθεί μέσα από το σύνολο των αναρτήσεων, συνοδευόμενα από κάποια στατιστικά και δείκτες. Κάθε θέμα περιέχει και υποκατηγορίες
14	Predicting the topic influence trends in social media with multiple models	2014	Όχι	Κοινωνικά δίκτυα	Ανίχνευση τάσεων	Ναι	Όχι	Συλλέγονται 18.012.123 χρήστες, οι σχέσεις μεταξύ τους και οι αναρτήσεις τους στο Twitter για περίοδο ενός μήνα (TW dataset). Για τον γεωγραφικό έλεγχο της μεθόδου χρησιμοποιείται ένα δεύτερο σύνολο δεδομένων από το κινεζικό δίκτυο Sina	Για δοσμένα από τον χρήστη θέματα, δίνεται η πρόβλεψη της μελλοντικής τους επιρροής με τη μορφή χρονοσειρών
15	Microblog Topic Contagiousness Measurement and Emerging Outbreak Monitoring	2014	Όχι	Twitter	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Όχι	Δϋο δύνολα δεδομένων από το Twitter, το πρώτο με 130 εκ. αναρτήσεις εκ των οποίων 3,3% με hashtags (11/2008-05/2009) και το δεύτερο με 79 εκ. αναρτήσεις εκ	Τα δημοφιλή θέματα τα οποία αντιπροσωπεύονται από το κεντρικό τους hashtag, και η παράμετρος διάδοσής τους

								των οποίων το 11% με hasgtags (01/2010-11/2010)	
16	Trend detection in social networks using Hawkes processes	2015	Όχι	Κοινωνικά δίκτυα	Ανίχνευση τάσεων	Όχι	Ναι	Δύο σύνολα δεδομένων: Το πρώτο έχει εξαχθεί από μέσο κοινωνικής δικτύωσης ενώ το δεύτερο από τις πιο ενεργές ιστοσελίδες για το διάστημα 03/2011-02/2012. Τα δεδομένα αυτά αντλήθηκαν από ιστοσελίδα πανεπιστημίου	Υπολογίζονται οι δείκτες της μεθοδολογίας για 10 προεπιλεγμένα θέματα και δίνονται τα αντίστοιχα γραφήματα (topic intensities, cumulative sum of jumps, maximum of CIR process).
17	Event detection and popularity prediction in microblogs	2015	Όχι	Microblogs	Ανίχνευση σημαντικών γεγονότων	Όχι	Ναι	Δύο σύνολα δεδομένων. Το πρώτο αποτελείται από 31 εκ. αναρτήσεις στο Twitter με παρακολούθηση τυχαίου δείγματος 313.000 χρηστών. Το δεύτερο από το κινέζικο μέσο Sina, για 119.000 χρήστες που σχετίζονται με τα δημοφιλή θέματα που προτείνει το μέσο	Σαν αποτελέσματα του πειράματος δίνονται τα σύνολα των ομαδοποιημένων δημοφιλών λέξεων (θέματα) και η πρόβλεψη της απήχησής τους.
18	Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data	2015	Ναι	Twitter	Ανίχνευση τάσεων	Όχι	Όχι	Γίνεται πειραματικός έλεγχος για 4 σύνολα δεδομένων που αφορούν 4 διαφορετικά θέματα προσδιορισμένα από πριν, με περίπου 40000 αναρτήσεις Twitter το κάθε ένα	Τα αποτελέσματα δίνονται στη χρήστη μέσα από μια διαδικτυακή πλατφόρμα με βασικές παροχές, και αφορούν την αναζήτηση που ο ίδιος έχει κάνει

19	An unsupervised framework of exploring events on twitter: filtering, extraction and categorization	2015	Όχι	Twitter	Ανίχνευση σημαντικών γεγονότων	Όχι	Όχι	Χρησιμοποιούνται δύο σύνολα δεδομένων με αναρτήσεις του Tweeter. Το πρώτο έχει συλλεχθεί χειροκίνητα και προσημανθεί και αφορά το Δεκέμβριο του 2010 ενώ το δεύτερο περιέχει 60 εκ. μη-προσημασμένες αναρτήσεις	Στην έξοδο του συστήματος περιλαμβάνονται τα πιο σημαντικά θέματα, χαρακτηριστικές λέξεις κλειδιά για το κάθε ένα καθώς και μια κατηγοριοποίηση των θεμάτων ανάλογα με το ευρύτερο πεδίο στο οποίο ανήκουν
20	Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks	2016	Όχι	Κοινωνικά δίκτυα	Ανίχνευση ανερχόμενων θεμάτων	Ναι	Ναι	Ένα σύνολο δεδομένων από το κοινωνικό δίκτυο Sina που αποτελείται από 13,973,119 tweets από 69, 394 χρήστες	Εντοπίζονται 54 ανερχόμενα και 50 μη-ανερχόμενα θέματα
21	Hot Topic Extraction Based on Chinese Microblog's Features Topic Model	2016	Όχι	Κοινωνικά δίκτυα	Ανίχνευση ανερχόμενων θεμάτων	Όχι	Ναι	Δεδομένα από το μέσο κοινωνικής δικτύωσης Sina. Το σύνολο περιέχει 200,000 αναρτήσεις που δημοσιεύτηκαν εκ των οποίων τα 2/3 χρησιμοποιούνται σαν δεδομένα μάθησης	Η έξοδος του συστήματος είναι τα δημοφιλή θέματα, και για τον έλεγχο της αποδοτικότητας ορίζεται ο δείκτης CR
22	Real-time event detection for online behavioral analysis of big social data	2017	Όχι	Κοινωνικά δίκτυα	Ανίχνευση σημαντικών γεγονότων	Ναι	Ναι	Πειραματική εφαρμογή σε δύο έτοιμα σύνολα δεδομένων από το Twitter από προηγούμενη μελέτη, τα "FA Cup" και "Super Tuesday"	Κάθε γεγονός αναπαρίσταται από ένα σύνολο λέξεων-κλειδιών και τα αποτελέσματα αξιολογούνται με βάση τρία χαρακτηριστικά: Επανάκληση θεματος (TOP-REC), ακρίβεια λέξης κλειδιού (K-PREC) και επανάκληση λέξης

									κλειδιού (K-REC), για κάθε θέμα

Πίνακας 6: Βιβλιογραφική επισκόπηση μεθοδολογιών ανίχνευσης τάσεων – Στάδια/Τεχνικές

	Δημοσίευση	**Τεχνικές/Τεχνολογίες**							
Α/Α	**Τίτλος**	**Στάδιο συλλογής δεδομένων**	**Στάδιο επεξεργασίας δεδομένων**		**Στάδιο ανίχνευσης θέματος**			**Στάδιο προσδιορισμού τάσης**	**Στάδιο ανάλυσης συναισθήματος**
			Γλωσσολογική επεξεργασία	**Άλλη επεξεργασία**	**Μοντέλο θέματος**	**Ομαδοποίηση**	**Άλλοι αλγόριθμοι**		
1	BlogPulse: Automated Trend Discovery for Weblogs	Αλγόριθμος ανίχνευσης ιστού στις πηγές (Intelliseek Spider).	Αφαίρεση κειμενων που δεν είναι γραμμένα στην αγγλική γλώσσα, τμηματοποίηση κειμένου, μετατροπή επιμέρους στοιχείων σε lower case	Τμηματοποίηση των HTML κόμβων	-	Ομαδοποίηση σε επίπεδο φράσης-κλειδιού, ομαδοποίηση βάσει ομοιότητας	-	Παρακολούθηση συχνότητας εμφάνισης θέματος στο χρόνο, γραφική παράσταση	-
2	TwitterStand: News in Tweets	Συλλογή των αναρτήσεων από τις διαφορετικές πηγές και φιλτράρισμα με τη βοήθεια ενός ενδιάμεσου σταδίου (classifier) για	Χρήση εργαλείων για επεξεργασία φυσικής γλώσσας (Natu-ral Language Processing-NLP)	Απόδοση ετικετών στις λέξεις, αναγνώριση επώνυμων οντοτήτων (με χρήση ειδικών εργαλείων)	-	Ομαδοποίηση βάσει γεωγραφικής τοποθεσίας (geographical clustering), ομαδοποίηση σε επίπεδο ανάρτησης	Gaussian, TF-IDF	TF-IDF, ανερχόμενα θέματα	-

		να αναγνωριστούν ως νέα ή όχι.							
3	Twittermonitor: trend detection over the twitter stream	Απευθείας από το Twitter API.	-	-	-	Ομαδοποίηση λέξεων-κλειδιών (GroupBurst), ομαδοποίηση βασισμένη σε ταυτόχρονη εμφάνιση	-	Ανερχόμενοι όροι (QueueBurst), ανερχόμενες λέξεις-κλειδιά	-
4	Monitoring trends on Facebook	Απλός αλγόριθμος που ανασύρει τις δημοσιες αναρτήσεις του Facebook ανά 10'.	Αφαίρεση stop-words, εξωτερικών συνδέσμων	-	-	Ομαδοποίηση βασισμένη σε ταυτόχρονη εμφάνιση, ομαδοποίηση βάσει κατανομής	-	Ανερχόμενοι όροι, TF-IDF	-
5	Social networking trends and dynamics detection via a cloud-based framework design (Cloud4Trends)	Αλγόριθμοι ανίχνευσης ιστού (Twitter, Blog parser).	Αφαίρεση stop-words, εξυγείανση κειμένου, stemming	Παρουσίαση δεδομένων από διαφορετικές πηγές με κοινό μοντέλο	-	Ομαδοποίηση βασισμένη σε χαρακτηριστικά, Γκαουσιανή ομαδοποίηση, ομαδοποίηση σε επίπεδο ανάρτησης	-	Ανερχόμενα θέματα, TF-IDF	-
6	Microblog Topic Detection Based on LDA Model and Single-Pass Clustering	Αλγόριθμος ανίχνευσης ιστού στις πηγές.	Σύστημα κατάτμησης κειμένου ICTCLAS (Segmantation system).	-	LDA	Ομαδοποίηση με ένα πέρασμα με χρήση του LDA, ομαδοποίηση σε επίπεδο ανάρτησης	-	-	-
7	Open Domain Event Extraction from Twitter (TwiCal)	Απευθείας από τη ροή αναρτήσεων του Twitter.	-	Απόδοση ετικετών στις λέξεις, αναγνώριση	-	LinkLDA, Bayesian μοντέλο για την ομαδοποίηση	Ανίχνευση φράσεων-θέματος με τη βοήθεια	Στατιστικός υπολογισμός παραμέτρων: Πιθανοτικός	-

				επώνυμων οντοτήτων		των γεγονότων κάτω από ευρύτερες κατηγορίες	λεξικολογικής ανάλυσης (POS tagger) και ενός εκπαιδευμένου μοντέλου (trained dataset)	υπολογισμός σημαντικότητας γεγονότος (αριθμός αναρτήσεων, χρονική σήμανση	
8	Deriving market intelligence from microblogs	Αλγόριθμος ανίχνευσης ιστού στις πηγές .	-	Απόδοση ετικετών στις λέξεις (Stanford POS Tagger)	-	-	-	Ανερχόμενοι όροι, TF-IDF	WordNet πλατφόρμα, SVM μοντέλο
9	Discovering hot topics using Twitter streaming data social topic detection and geographic clustering	Απευθείας από το Twitter API.	-	Απόδοση γεωγραφικών ετικετών στις αναρτήσεις (GEO tagging), εξαίρεση αναρτήσεων που δεν είναι μέσα σε ένα γεωγραφικό εύρος	-	Ομαδοποίηση με χειροκίνητο τρόπο (οπτική επισκόπηση των λέξεων-κλειδιών)	-	Ανερχόμενοι όροι, ανερχόμενες λέξεις-κλειδιά, TF, ομαδοποίηση ανά περιοχή	-
10	Detecting trends on the Web: A multidisciplinary approach	Αλγόριθμος ανίχνευσης ιστού στις πηγές, αλγόριθμος ανίχνευσης άλλων πιθανών πηγών .	Αφαίρεση stop-words, html στοιχείων, εξωτερικών συνδέσμων, stemming	-	LDA	-	-	Χειροκίνητος προσδιορισμός τάσης	SentiWordNet πλατφόρμα
11	Hashtag Graph based Topic Model for Tweet Mining	-	Βασικά βήματα κανονικοποίησης, εξαίρεση αναρτήσεων που δεν περιέχουν hashtags	Εξαίρεση αναδημοσιεύσεων (retweets)	HGTM	Ομαδοποίηση βάσει των hashtags	-	-	-

12	PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis	Αλγόριθμος ανίχνευσης ιστού στις πηγές .	-	Εξαγωγή hashtags, εξαγωγή hashtag συναισθήματος , εξαγωγή εικονιδίων	-	Ομαδοποίηση βάσει των hashtags	-	Παρακολούθηση συχνότητας εμφάνισης θέματος στο χρόνο (αριθμός αναρτήσεων ανά περίοδο)	Χρήση των hastag συναισθήματος σε σύγκριση με την προηγούμενη περίοδο
13	A prerecognition model for hot topic discovery based on microblogging data	Αλγόριθμος ανίχνευσης ιστού στις πηγές (microblog APIs)	Αφαίρεση stop-words, αφαίρεση άχρηστων λέξεων	Εξαγωγή hashtags	LDA, PAM	Ομαδοποίηση λέξεων-κλειδιών, ομαδοποίηση θεμάτων	Αναγωγή hashtags απευθείας σε θέματα, k-NN αλγόριθμος	Στατιστικός υπολογισμός παραμέτρων	-
14	Predicting the topic influence trends in social media with multiple models	-	Αφαίρεση stop-words, stemming	-	-	-	-	Αλγόριθμος K-NN	-
15	Microblog Topic Contagiousness Measurement and Emerging Outbreak Monitoring	-	Εξαίρεση των αναρτήσεων που δεν περιέχουν hashtags	-	PLDA	-	Stanford Topic Modeling Toolbox	Στατιστικός υπολογισμός παραμέτρων: Υπολογισμός παραμέτρου διάδοσης R με χρήση Bayesian-inference Python package, μοντέλο SIR	-
16	Trend detection in social networks using	-	-	-	-	Χειροκίνητη ομαδοποίηση	-	Hawkes process,	-

	Hawkes processes							στοχαστικό μοντέλο	
17	Event detection and popularity prediction in microblogs	-	Αφαίρεση stop-words, εξαίρεση αναρτήσεων με λιγότερες από 3 λέξεις	-	-	Ομαδοποίηση βασισμένη σε ταυτόχρονη εμφάνιση, ομαδοποίηση βάσει κατανομής (clustering by distribution)	Αλγόριθμος Viterbi	Ανερχόμενες λέξεις-κλειδιά, απόδοση βάρους στις λέξεις, LDA για την πρόβλεψη ενδιαφέροντος των χρηστών, γραμμικό μοντέλο διάδοσης για το θέμα	-
18	Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data	Γίνεται με τη βοήθεια του εργαλείου Sysomos, το οποίο εμφανίζει και τη χρονική σφραγίδα, την τοποθεσία, το συναίσθημα και την ταυτότητα του χρήστη	Φιλτράρισμα stop-words, εξαίρεση λέξεων ξένης γλώσσας, εξαίρεση αναρτήσεων με πολλά σύμβολα.	Χρονική ταξινόμηση και καταμέτρηση αναρτήσεων σε ημερήσια βάση	-	-	CMU tweet parser, POS tagger	Χρήση οντολογιών, μέτρηση συχνότητας εμφάνισης	-
19	An unsupervised framework of exploring events on twitter: filtering, extraction and categorization	Απευθείας από τη ροή αναρτήσεων του Twitter	Stemming	Απόδοση ετικετών στις λέξεις, αναγνώριση επώνυμων οντοτήτων	-	LECM μοντέλο (ανίχνευση γεγονότων και κατηγοριοποίησή τους), ομαδοποίηση σε επίπεδο θέματος (Bayesian μοντέλο)	Φιλτράρισμα των αναρτήσεων ανάλογα με το αν περιέχουν λέξεις μέσα από ένα συγκεκριμένο σύνολο, φιλτράρισμα βάσει	-	-

							χαρακτηριστικών, SVM μοντέλο		
20	Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks	Απευθείας από το Sina API	-	-	-	Ομαδοποίηση λέξεων-κλειδιών (DBSCAN)	K-Means	Ανερχόμενοι όροι (DBN), ανερχόμενες λέξεις-κλειδιά,TF, PageRank	-
21	Hot Topic Extraction Based on Chinese Microblog's Features Topic Model	-	Αφαίρεση stop-words, αφαίρεση στίξης	Ομαδοποίηση αναρτήσεων κατά ημερομηνία	MF-LDA	Ομαδοποίηση βασισμένη σε χαρακτηριστικά	-	Ανερχόμενοι όροι, ανερχόμενες αναρτήσεις	-
22	Real-time event detection for online behavioral analysis of big social data	Απευθείας από το Twitter API με όρο αναζήτησης	Εξυγείανση κειμένου,αφαίρεση συμβόλων, συνδέσμων, ετικετών, αφαίρεση περιττών λέξεων	Εισαγωγή δεδομένων σε κανονικοποιημένη μορφή σε βάση, αναγωγή hashtags σε λέξεις-κλειδιά	-	Ομαδοποίηση αναρτήσεων, ομαδοποίηση βάσει της σημασιολογικής απόστασης (γράφος), OPTICS αλγόριθμος	-	Ανερχόμενοι όροι	-

Πίνακας 7: Βιβλιογραφική επισκόπηση μεθοδολογιών ανίχνευσης τάσεων – Βήματα

	Δημοσίευση	**Βήματα μεθοδολογίας**
1	Detecting trends on the Web: A multidisciplinary approach [23]	1. Ένας απλός crawling αλγόριθμος εντοπίζει όλα τα αντικειμενικά (factual) κείμενα σε ένα σύνολο πηγών (document retrieval) 2. Πριν την εφαρμογή των αλγορίθμων εύρεσης θέματος και ανάλυσης συναισθήματος πραγματοποιείται επεξεργασία των αρχείων με αφαίρεση των περιττών λέξεων, των εξωτερικών συνδέσμων κ.ο.κ. 3. Ένα LDA μοντέλο θέματος εξάγει τις θεματικές ενότητες σε κάθε περίοδο από τα αντικειμενικά κείμενα, αφού έχει εκπαιδευτεί σύμφωνα με τα δεδομένα των 2 προηγούμενων περιόδων 4. Με χρήση 2 αλγορίθμων γίνεται επέκταση στο σετ των πηγών ανάλογα με τις αναφορές που έχουν στα κείμενα που έχουν ήδη μελετηθεί 5. Ένας λεξικολογικός (lexicon-based) αλγόριθμος βασισμένος στο SentiWordNet (resource of lexical information) χρησιμοποιείται για να κάνει ανάλυση συναισθήματος στα υποκειμενικά κείμενα (opinionated documents), δηλαδή στις αναρτήσεις των κοινωνικών δικτύων που αναφέρονται στα θέματα που έχουν εντοπιστεί 6. Οι τάσεις (δημοφιλή θέματα) επιλέγονται χειροκίνητα από τα θέματα που έχουν εντοπιστεί
2	Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks [30]	1. Τα δεδομένα συλλέγονται απευθείας από το API του Sina. Δεν μεσολαβεί επεξεργασία δεδομένων 2. Επιλέγονται δύο χαρακτηριστικά (features) τα οποία διαχωρίζουν ένα ανερχόμενο από ένα μη-ανερχόμενο θέμα: ελκυστικότητα (attractiveness) που ορίζει αν το θέμα έλκει τους χρήστες να το διαδώσουν) και κόμβοι-κλειδιά (key-node): ορίζει αν υπάρχουν χρήστες με επιρροή οι οποίοι μπορούν να προκαλέσουν μεγάλο αριθμό αναμεταδόσεων και να διαδώσουν σημαντικά το θέμα 3. Δημιουργείται ένα DBN μοντέλο (probabilistic graphical model) με δύο κρυφές μεταβλητές οι οποίες αντιστοιχούν στα χαρακτηριστικά που έχουν επιλεχθεί και τέσσερις μεταβλητές παρατήρησης (αριθμός κόμβων, αριθμός retweeting chains, μέγιστος αριθμός κόμβων στις αλυσίδες και ο συνολικός αριθμός των followers) . Κάθε μεταβλητή εντάσσεται σε 3 καταστάσεις με χρήση του K-means αλγορίθμου. 4. Το μοντέλο εκπαιδεύεται (παραμετροποιείται) με βάση μια ακολουθία training data και στη συνέχεια υπολογίζει τις πιθανότητες των κρυφών μεταβλητών για τα πραγματικά δεδομένα. Το τελικό αποτέλεσμα είναι η πιθανότητα ένα keyword να είναι emerging ή όχι στη συγκεκριμένη χρονική στιγμή. Στον εντοπισμό των ανερχόμενων λέξεων-κλειδιών χρησιμοποιείται επιλογή χαρακτηριστικών (feature selection).

		5. Αφού έχει εξαχθεί μια λίστα με emerging keywords, αναλύεται η αλληλοσυσχέτιση μεταξύ τους ώστε να ομαδοποιηθούν σε emerging topics για τη συγκεκριμένη χρονική στιγμή. Αυτό επιτυγχάνεται με χρήση του DBSCAN αλγορίθμου. Η ομαδοποίηση μπορεί να είναι είτε **επιβλεπόμενη** (supervised) είτε **μη-επιβλεπόμενη** (unsupervised).
3	Trend detection in social networks using Hawkes processes [24]	1. Δεν περιλαμβάνεται στάδιο συλλογής δεδομένων ούτε στάδιο επεξεργασίας τους, τα πειράματα γίνοντα σε έτοιμα σύνολα 2. Το μέσο κοινωνικής δικτύωσης αναπαρίσταται σαν ένα κατευθυνόμενος γράφος G(V,E), όπου V το σύνολο των χρηστών και E οι μεταξύ τους συνδέσεις. Σε μία περίοδο οι χρήστες πραγματοποιούν αναρτήσεις σχετικά με K προκαθορισμένα θέματα (δεν περιλαμβάνεται στάδιο ανίχνευσης θέματος). Οι συγγραφείς ξεκαθαρίζουν ότι τα θέματα έχουν εντοπιστεί εκ των προτέρων μέσω τεχνικών εξόρυξης πληροφορίας (text mining) οι οποίες δεν αναφέρονται στη μελέτη και δεν αποτελούν αντικείμενο αυτής. Έτσι κάθε ανάρτηση χρήστη αφορά ένα μόνο προκαθορισμένο θέμα K, και το αντικείμενο της μελέτης είναι να προσδιοριστεί αν το K είναι ανερχόμενο θέμα ή όχι. 3. Χρησιμοποιείται η στοχαστική διασικασία Hawkes (linear Hawkes process) X, όπου το στοιχείο X(i,k,t) αναπαριστά τον αριθμό των αναρτήσεων για το θέμα k που κάνει ο χρήστης i την περίοδο t. Η διαδικασία Hawkes είναι μια στοχαστική διαδικασία σημείου (point process) που χρησιμοποιείται για προσδιορισμό μοτίβων σε ένα σύνολο σημείων. 4. Στη συνέχεια ο αλγόριθμος αναζήτησης των τάσεων, σύμφωνα με τη θεωρία της σχεδόν ασταθούς διαδικασίας Hawkes ψάχνει το αναμενόμενο μέγιστο κατάλληλων δεικτών για το κάθε θέμα k. Οι δείκτες είναι χρονοεξαρτώμενοι και καταγράφουν τις κορυφώσεις κάθε θέματος κατά τη διάρκεια διάδοσης. Συγκεκριμένα υπολογίζεται η μέγιστη αναμενόμενη ένταση διάδοσης (maximum expected broadcast intensity) λαμβάνοντας υπόψιν και την τοπολογία του δικτύου. 5. Στο τέλος ο αλγόριθμος υπολογίζει τους δείκτες τάσης για κάθε θέμα σαν ολοκλήρωμα των προηγούμενων.
4	Twittermonitor: trend detection over the twitter stream [31]	1. Τα δεδομένα παρέχονται απευθείας από το API του Twitter. Στο back end της εφαρμογής γίνεται δειγματοληψία, ώστε να μειώνεται ο όγκος των δεδομένων προς επεξεργασία από 50 εκ. Στα 10 εκ. Αναρτήσεις ανά ημέρα. Το στάδιο αναφέρεται ως streamlistener.

		2. Το στάδιο ανίχνευσης τάσης εντοπίζει τις λέξεις-κλειδιά που εμφανίζονται σε πολλές αναρτήσεις (tweets) κάνοντας χρήση του αλγορίθμου QueueBurst. Στον εντοπισμό των ανερχόμενων λέξεων-κλειδιών χρησιμοποιείται μόνο το κριτήριο της συχνότητας με την οποία αναφέρονται σε πραγματικό χρόνο. 3. Το στάδιο ανίχνευσης θέματος τις ομαδοποιεί σε σύνολα, βασιζόμενο στην συχνότητα με την οποία εμφανίζονται μαζί (Clustering based on co-occurences, Unsupervised Topic Detection) με τον αλγόριθμο GroupBurst. Έτσι η τάση ορίζεται σαν ένα σύνολο από δημοφιλείς λέξεις-κλειδιά οι οποίες εμφανίζονται συχνά μαζί στις αναρτήσεις. 4. Γίνεται περαιτέρω ανάλυση στις αναρτήσεις για να εξαχθούν χρήσιμες πληροφορίες σχετικά με την τάση. Εντοπίζονται περισσότερες σχετικές λέξεις-κλειδιά (αλγόριθμοι context extraction όπως PCA, SVD), εντοπίζεται η γεωγραφική περιοχή όπου υφίσταται η ταση και τέλος στην περιγραφή προστίθενται πηγές από μεγάλες ιστοσελίδες οι οποίες αναφέρονται συχνά. Τα στοιχεία των τάσεων ανανεώνονται σε πραγματικό χρόνο οσο αυτό παραμένει δημοφιλές. 5. Τα αποτελέσματα προβάλλονται στον τελικό χρήστη μέσω ενός μιας ολοκληρωμένης διαδικτυακής πλατφόρμας (web interface). **-Πρόκειται για ολοκληρωμένη, υλοποιημένη πλατφόρμα-εργαλείο που απευθύνεται σε τελικούς χρήστες.**
5	Monitoring trends on Facebook [17]	1. Η συλλογή των δεδομένων είναι συνεχόμενη, για να εξάγονται συμπεράσματα σε πραγματικό χρόνο, και αφορά το πλήρες σύνολο των δημοσίων αναρτήσεων. Πραγματοποιείται με έναν απλό αλγόριθμο πάνω στο Facebook Graph API. Ο αλγόριθμος εκτελείται και συγκεντρώνει αναρτήσεις (Fecebook posts) κάθε 10 λεπτά. 2. Πριν την ανίχνευση γίνεται μια γλωσσολογική επεξεργασία στις αναρτήσεις (stop-word filtering, url removal). . Αφαιρούνται οι λέξεις που δεν έχουν σημασιολογικό περιεχόμενο (stop-words) σύμφωνα με μια προκαθορισμένη λίστα καθώς και οι σύνδεσμοι (url). 3. Για τον προσδιορισμό του θέματος (post topic identification) χρησιμοποιείται η προσέγγιση TF-IDF, Term Frequency-Inverse Document Frequency όπου οι όροι βαθμονομούνται ανάλογα με τη συχνότητά τους στις αναρτήσεις. 4. Στην ανίχνευση ομάδων (cluster detection), η ομαδοποίηση γίνεται με δύο διαφορετικούς αλγορίθμους, έναν με βάση την κατανομή των όρων και έναν με βάση της συχνότητας με την οποία εμφανίζονται μαζί σε μια ανάρτηση (co-occurence).

		5. Γίνεται ομαδοποίηση δημοφιλών όρων (terms) οι οποίοι αποτελούνται κατ'ελάχιστο από 2 λέξεις ο καθένας. Για την ανίχνευση θέματος χρησιμοποιείται η τεχνική TF-IDF η οποία καταλήγει σε μια λίστα με τους δημοφιλείς όρους. Στη συνέχεια οι όροι αυτοί ομαδοποιούνται με βάση την κατανομή τους και με τη συχνότητα εμφάνισής τους σε δημοφιλή θέματα (Clustering based on co-occurences, Clustering based on distribution).
6	Cloud4Trends [26]	1. Συλλογή δεδομένων από τη ροή του Twitter καθώς και από ένα σύνολο επιλεγμένων blogs, που αφορούν έναν αριθμό γεωγραφικών περιοχών. Τα δεδομένα αντλούνται από το Twitter Streaming API και από το Google Blogger API σε πραγματικό χρόνο, μέσω parsers. 2. Εφαρμογή μιας τεχνικής ομαδοποίησης στα δεδομένα που έχουν συλλεχθεί για να βρεθούν τα πρόσφατα δημοφιλή θέματα. Τα δεδομένα από τις διαφορετικές πηγές αναπαρίστανται με ενιαίο μοντέλο, το οποίο συμπεριλαμβάνει και ένα TF-IDF key value map, που χρησιμεύει αργότερα για την ομαδοποίηση. Στο στάδιο εμπεριέχεται προ-επεξεργασία προκειμένου να αφαιρεθούν οι περιττές λέξεις (stop-words) και να μείνουν μόνο τα χρήσιμα χαρακτηριστικά του κειμένου. 3. Η ομαδοποίηση με βάση το θέμα γίνεται λαμβάνοντας υπόψιν συγκεκριμένα χαρακτηριστικά (feature/attribute based) και σε επίπεδο ανάρτησης (document based) και όχι λέξεως-κλειδιού. Οι θεματικές ομάδες που έχουν εντοπιστεί χαρακτηρίζονται ως ενεργές και ανενεργές. Οι ενεργές αξιολογούνται σύμφωνα με μια κλίμακα τάσης και αυτές με το υψηλότερο σκορ παρουσιάζονται ως τάσεις στον τελικό χρήστη. 4. Εκκαθάριση και βαθμονόμηση των ομάδων έτσι ώστε οι τάσεις να προσδιοριστούν επακριβώς και να οπτικοποιηθούν.
7	Hot Topic Extraction Based on Chinese Microblog's Features Topic Model [32]	1. Η μεθοδολογία εφαρμόζεται σε έτοιμα σύνολα δεδομένων και δεν περιλαμβάνει στάδιο συλλογής και επεξεργασίας δεδομένων 2. Ορίζονται 3 τιμές που υπολογίζονται με βάση 5 χαρακτηριστικά του μέσου δικτύωσης (microblog). Αποτίμηση προσοχής (Attention value, υπολογίζεται από τα χαρακτηριστικά support,retweet,comment), Αποτίμηση επιρροής (Authority value, υπολογίζεται από τον αριθμό των χρηστών με μεγάλη επιρροή που αναδημοσιεύουν την ανάρτηση με τη βοήθεια PageRank αλγορίθμου) και Συχνότητα λέξης (Word Frequency, υπολογίζεται από τη χρονική σήμανση) και τη χρονική θυρίδα όπου γίνεται η επεξεργασία.

		3. Πραγματοποιείται μια προ-επεξεργασία γλωσσολογικού χαρακτήρα στα δεδομένα του κοινωνικού δικτύου και ομαδοποιούνται με βάση το χρόνο. Γίνεται ομαδοποίηση ανάλογα με την ημερομηνία δημοσίευσης, εφαρμογή ενός εργαλείου για την αντιμετώπιση των λέξεων που έχουν σπάσει σε τμήματα (segmentation) και φίλτρο για να αφαιρεθούν οι περιτές λέξεις. 4. Υπολογίζεται η πρώτη τιμή (attention value) για την κάθε ανάρτηση και απορρίπτονται όσες δεν φτάνουν μέχρι ένα ορισμένο κατώφλι. 5. Οι δύο υπολειπόμενες τιμές υπολογίζονται και λαμβάνονται υπόψιν στην εφαρμογή του LDA μοντέλου. Γίνεται με συνδυασμό πιθανοτικού μοντέλου θέματος (LDA) και ομαδοποίησης βασισμένης σε χαρακτηριστικά (Feature-based). Η υβριδική μέθοδος ονομάστηκε MF-LDA.
8	Event detection and popularity prediction in microblogs [38]	1. Η συλλογή των δεδομένων γίνεται απευθείας από τα API των Twitter και Sina 2. Για κάθε ανάρτηση (micro-blog document) διαχωρίζονται οι λέξεις και αφαιρούνται οι λέξεις χωρίς σημασιολογικό περιεχόμενο. Επίσης εξαιρούνται από το δείγμα οι αναρτήσεις με λιγότερες από 3 λέξεις 3. Έπειτα γίνεται συνδυασμός των σχέσεων μεταξύ των χρηστών και της κατανομής της συχνότητας με την οποία εμφανίζονται οι λέξεις για να εντοπιστούν οι δημοφιλείς όροι (burst words). Αυτό γίνεται με απόδοση βάρους στις λέξεις των αναρτήσεων (word weighting) , ανίχνευση δημοφιλών λέξεων και ομαδοποίηση δημοφιλών λέξεων σε θέματα. Τα πρώτα δύο βήματα χρησιμοποιούν πιθανοτικά μοντέλα ενώ το τρίτο βασίζεται στη δημιουργία ενός γράφου συσχέτισης των λέξεων. 4. Οι δημοφιλείς όροι ομαδοποιούνται, και κάθε μία από τις ομάδες μπορεί να θεωρηθεί ότι σχετίζεται με ένα συγκεκριμένο θέμα/γεγονός. Για την ομαδοποίηση, χρησιμοποιείται ένας γράφος συσχέτισης λέξεων σύμφωνα με το πόσο συχνά εμφανίζονται μαζί στα μέσα κοινωνικής δικτύωσης και σε άλλες πηγές. Κάθε κόμβος είναι μια δημοφιλής λέξη και κάθε ακμή αναπαριστά την σχέση μεταξύ δύο λέξεων. Σε αυτόν δημιουργούνται υπογράφοι στενά συνδεδεμένων λέξεων οι οποίοι είναι τα θέματα. Η εγγύτητα των λέξεων-κλειδιών υπολογίζεται με βάση τη συχνότητα εμφάνισής τους μαζί στην ίδια ανάρτηση και τις κατανομές τους 5. Γίνεται μοντελοποίηση της διάδοσης ενός γεγονότος κάνοντας χρήση των αναρτήσεων που το αφορούν και των σχέσεων μεταξύ των χρηστών στο κοινωνικό δίκτυο. Το μοντέλο χρησιμοποιεί μια γραμμική συνάρτηση η οποία εμπεριέχει όλες τις σχετικές πληροφορίες (επιρροή, ενδιαφέρον χρήστη, ιστορικά στοιχεία). Η δημοτικότητα του γεγονότος προκύπτει ως αποτέλεσμα της συνάρτησης για όλους τους χρήστες. Για το ενδιαφέρον του χρήστη χρησιμοποιείται ένα LDA μοντέλο θέματος.

9	Deriving market intelligence from microblogs [27]	1. Τα δεδομένα από το μέσο κοινωνικής δικτύωσης όπου γίνεται το πείραμα (απόψεις και σχέσεις μεταξύ των χρηστών) συλλέγονται με τη βοήθεια ενός αλγορίθμου προσπέλασης (web spider). 2. Ένα αντίγραφο των δεδομένων προσημαίνεται (POS-tagged) με ετικέτες προκειμένου να διευκολυνθεί η μετέπειτα σημασιολογική ανάλυση. Δημιουργείται γράφος των χρηστών οι οποίοι σχετίζονται με τις αναρτήσεις ώστε να γίνει αργότερα η αξιολόγηση αξιοπιστίας. 3. Ανίχνευση θέματος: χρησιμοποιείται ο όρος αναζήτηση (query) για να προσδιοριστεί η θεματική ενότητα για την οποία ο ενδιαφερόμενος θέλει να αντλήσει πληροφορία (για παράδειγμα το όνομα μιας εταιρείας ή ένα προϊόν. Κάθε δημοφιλής όρος που εντοπίζεται βαθμονομείται ανάλογα με το πόσο πιθανό είναι να σχετίζεται με κάποιο θέμα που αφορά τους όρους της αναζήτησης (query) (TF-IDF). Έτσι το σύστημα (trendy topic detection module) θα εντοπίσει τα σχετικά με την αναζήτηση ανερχόμενα θέματα (trendy topics) για να προχωρήσει στην ανάλυση. 4. Ταξινόμηση των απόψεων: Την ταξινόμηση (classification) των απόψεων πραγματοποεί ένα SVM module. Γίνεται χρήση του διαθέσιμου WordNet σημασιολογικού λεξικού και στη συνέχεια χρησιμοποιείται το μοντέλο SVM που προσδιορίζει τον συναισθηματικό προσανατολισμό (sentiment polarity) των αναρτήσεων. 5. Αξιολόγηση Αξιοπιστίας (Credibility assessment) 6. Ενοποίηση των αποτελεσμάτων και δημιουργία αναφοράς (report). Στο τελευταίο στάδιο και αφού ληφθεί υπόψιν και η αξιοπιστία των δεδομένων δίνεται ένα σκορ για το κάθε σχετικό θέμα.
10	Hashtag Graph based Topic Model for Tweet Mining [53]	1. Δεν υπάρχει στάδιο συλλογής δεδομένων. Στον πειραματικό έλεγχο της μεθοδολογίας χρησιμοποιείται ένα έτοιμο σύνολο δεδομένων 2. Πραγματοποιείται μια στοιχειώδης επεξεργασία και φιλτράρισμα ώστε να μείνουν οι πιο ποιοτικές αναρτήσεις του συνόλου. 3. Δημιουργείται γράφος για το σύνολο των δεδομένων, με κόμβους τα hashtags και ακμές που βαθμονομούνται ανάλογα με τη συσχέτιση των κόμβων μεταξύ τους. Αντίστοιχα οι σχέσεις μπορούν να αναπαρασταθούν και σε μητρική μορφή. 4. Πραγματοποιείται μια στοιχειώδης επεξεργασία και φιλτράρισμα ώστε να μείνουν οι πιο ποιοτικές αναρτήσεις του συνόλου. Η μεθοδολογία παρέχει ένα εναλλακτικό πιθανοτικό μοντέλο το οποίο ανιχνεύει θέματα χρησιμοποιώντας την τεχνική της ομαδοποίησης σε επίπεδο λέξεων και σε επίπεδο hashtags.

		5. Παρόμοια με το LDA μοντέλο, κάθε θέμα αναπαρίσταται σαν μια πιθανοτική κατανομή (distribution) από λέξεις.Κάθε hashtag αναπαρίσταται σαν μια κατανομή θεμάτων. 6. Για κάθε νέο hashtag συμπεραίνεται η κατανομή του στα θέματα.
11	PoliTwi: Early detection of emerging political topics on twitter [33]	1. Τα tweets συλλέγονται απευθείας από το API του Twitter με τη βοήθεια ενός crawler και αποθηκεύονται σε μια βάση δεδομένων PostgreSQL. Οι όροι αναζήτησης προσδιορίζονται από το χρήστη. 2. Ακολουθεί η προεπεξεργασία των αναρτήσεων. Εξάγονται τα hashtags που θα χρησιμοποιηθούν για την ανίχνευση του θέματος, σημειώνονται τα «hashtag συναισθήματος» που θα χρησιμοποιηθούν για την ανάλυση συναισθήματος (το ίδιο και τα emoticons που υπάρχουν στις αναρτήσεις) και τα επεξεργασμένα δεδομένα αποθηκεύονται σε νέα βάση. 3. Στο στάδιο της ανάλυσης γίνεται η ανίχνευση των κορυφαίων θεμάτων, τα οποία αντιπροσωπεύονται από τα αντίστοιχα hashtags. 4. Ακολουθεί η παρουσίαση των αποτελεσμάτων στους χρήστες μέσα από τρία διαφορετικά κανάλια (twitter, website, mobile app).
12	A prerecognition model for hot topic discovery based on microblogging data [34]	1. Γίνεται συλλογή δεδομένων από το API του μέσου κοινωνικής δικτύωσης/microblog με τη βοήθεια ενός crawler. 2. Φιλτράρισμα ή προ-επεξεργασία των αναρτήσεων για να αφαιρεθούν οι ασήμαντες λέξεις, ενοποίηση των σημασιολογικά σημαντικών λέξεων σε λέξεις κλειδιά και απευθείας αναγωγή των hashtags σε θέματα. 3. Γίνεται ομαδοποίηση στις αρχικές αναρτήσεις για να εξαχθούν τα θέματα και η «ποσότητα» των θεμάτων. Η διαδικασία αυτή χωρίζεται σε επιμέρους στάδια. Αρχικά χρήση των LDA και PAM μοντέλων θέματος για να εξαχθούν τα θέματα και τα υπο-θέματα και τέλος ομαδοποίηση των συγγενών θεμάτων με χρήση του KNN (K-Nearest Neighbor) αλγορίθμου. 4. Στατιστικός υπολογισμός παραμέτρων και έλεγχος κατωφλίων για κάθε θέμα που έχει ανιχνευθεί: Υπολογισμός της εμβέλειας και της επιτάχυνσης (όπως έχουν οριστεί στα πλαίσια της μεθοδολογίας) του κάθε θέματος. Ορίζονται σημεία μεταβολής ανάλογα με τη διάρκεια ζωής (lifecycle) του θέματος και αντίστοιχα κατώφλια.

		5. Επιλέγονται πιθανά δημοφιλή θέματα στην κατάλληλη περίοδο του κύκλου ζωής τους.
13	Discovering hot topics using Twitter streaming data social topic detection and geographic clustering [35]	1. Συλλογή των αναρτήσεων με την επιθυμητή γεωγραφική σήμανση (geo-tag) σε συγκεκριμένα όρια, απευθείας από το Twitter API. Γίνεται περιορισμός στα γεωγραφικά προσημασμένα δεδομένα για ένα συγκεκριμένο εύρος 2. Υπολογισμός της συχνότητας εμφάνισης των λέξεων και εντοπισμός των ανερχόμενων λέξεων-κλειδιών. 3. Η ομαδοποίηση των δημοφιλών λέξεων σε θέματα γίνεται με χειροκίνητο τρόπο (οπτική επισκόπηση των λέξεων-κλειδιών) 4. Για κάθε θέμα που έχει ανιχνευθεί γίνεται ομαδοποίηση ανά γεωγραφική περιοχή και ανάλυση.
14	Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data [43]	1. Η συλλογή δεδομένων γίνεται με τη βοήθεια του εργαλείου Sysomos, το οποίο εμφανίζει και τη χρονική σφραγίδα, την τοποθεσία, το συναίσθημα και την ταυτότητα του χρήστη. 2. Στο στάδιο επεξεργασίας των δεδομένων έχουμε φιλτράρισμα stop-words, εξαίρεση λέξεων ξένης γλώσσας, εξαίρεση αναρτήσεων με πολλά σύμβολα. Χρονική ταξινόμηση και καταμέτρηση των αναρτήσεων σε ημερήσια βάση. Εξάγονται οι σημασιολογικές οντότητες των αναρτήσεων. 3. Για την ανίχνευση τάσεων χρησιμοποιούνται οντολογίες και μετράται η συχνότητα εμφάνισής τους.
15	Predicting the topic influence trends in social media with multiple models [29]	1. Η εργασία γίνεται σε έτοιμο σύνολο δεδομένων, δεν προηγείται συλλογή. 2. Πραγματοποιείται αφαίρεση άχρηστων λέξεων και stemming. 3. Γίνεται η υπόθεση ότι τα θέματα έχουν ανιχνευθεί σε προηγούμενα στάδια και δίνονται στο σύστημα με τη μορφή λέξεων-κλειδιών, άρα δεν έχουμε ανίχνευση θέματος. 4. Στο σύνολο δεδομένων μετράται η συχνότητα εμφάνισης του θέματος και η δημοτικότητά του, με πιθανοτικό τρόπο μέσω συναρτήσεων. Γίνεται χρήση μεθόδων πάνω σε γράφους για να ληφθεί υπόψιν και ο γεωγραφικός παράγοντας διάδοσης (k-NN αλγόριθμος). 5. Για δοσμένα από τον χρήστη θέματα, δίνεται η πρόβλεψη της μελλοντικής τους επιρροής με τη μορφή χρονοσειρών.
16	BlogPulse: Automated Trend Discovery for Weblogs [36]	1. Για να συλλεχθούν οι πηγές, αρχικά συγκεντρώθηκε μια λίστα 22,000+ weblogs τον Φεβρουάριο του 2003, η οποία ανανεωνόταν με αυτόματο τρόπο έως ότου οι πηγές έφτασαν το μέγιστο που μπορούσε να διαχειριστεί το σύστημα (100.000).

		2. Γίνεται προσπέλαση των πηγών με τη βοήθεια του Intelliseek Spider. Η διαδικασία αυτή διαρκει περίπου 12 ώρες. 3. Γίνεται αφαίρεση κειμενων που δεν είναι γραμμένα στην αγγλική γλώσσα, τμηματοποίηση κειμένου, μετατροπή επιμέρους στοιχείων σε lower case, τμηματοποίηση των HTML κόμβων. 4. Το BlogPulse δημιουργεί καθημερινά ένα ενιαίο σώμα κειμένου από τη συλλογή των αναρτήσεων στα blogs, χρησιμοποιώντας ένα πακέτο εφαρμογών που πραγματοποιεί εύρεση φράσεων, ευρετηριοποίηση, ανάλυση τάσεων και εξόρυξη δεδομένων (Analyst Workbench). 5. Στη συνέχεια αφού έχει γίνει η εύρεση των φράσεων-κλειδιών, υπολογίζεται συχνότητα εμφάνισής τους και ομαδοποιούνται σε θέματα ανάλογα με την ομοιότητά τους. 6. Η ανάλυση τάσεων γίνεται (πχ για ένα προιόν) ανάλογα με τη συχνότητα με την οποία αναφέραται στο σώμα του κειμένου που αναλύεται. Αυτό παρακολουθείται για ένα συνεχόμενο χρονικό διάστημα και παράγεται η γραφική παράσταση της τάσης.
17	Microblog Topic Detection Based on LDA Model and Single-Pass Clustering [28]	1. Εφαρμογή ενός crawler αλγορίθμου στις πηγές (microblogs) για τη συλλογή των δεδομένων. 2. Προεπεξεργασία των δεδομένων κάνοντας χρήση του συστήματος κατάτμησης κειμένου ICTCLAS (Segmantation system). 3. Ομαδοποίηση αναρτήσεων σε θέματα με ένα πέρασμα (single pass clustering) με τη βοήθεια του LDA μοντέλου. 4. Παρουσίαση αποτελεσμάτων και σύγκριση με το VSM μοντέλο.
18	TwitterStand: News in Tweets [39]	1. Συλλογή των αναρτήσεων από τις διαφορετικές ροές και φιλτράρισμα με τη βοήθεια ενός ενδιάμεσου σταδίου (classifier) για προσδιοριστεί αν αφορούν νέα ή όχι. Στόχος είναι να εξαλειφθεί τελείως ο «θόρυβος». Οι διαφορετικές ροές είναι 4: 2000 επιλεγμένοοι χρήστες που δημοσιεύουν συχνά γεγονότα, όπως τηλεοπτικοί σταθμοί (Seeders), μια ροή αναρτήσεων που αποτελεί δείγμα από όλες τις δημοσιεύσεις όλων των χρηστών (GardenHose), μια υπηρεσία που έχει διαμορφωθεί για αναζήτηση αναρτήσεων με συγκεκριμένες λέξεις-κλειδιά (Search) και μια αλγοριθμική επιλογή χρηστών (BirdDog) που προστίθενται σταδιακά

		2. Χρήση εργαλείων για επεξεργασία φυσικής γλώσσας και απόδοση ετικετών (Natu- ral Language Processing-NLP, Part-Of-Speech-POS tagging). Χρήση εργαλείων για αναγνώριση επώνυμων οντοτήτων (Named-Entity Recognition-NER). 3. Στο στάδιο της ανίχνευσης θέματος, οι αναρτήσεις ομαδοποιούνται ώστε κάθε ομάδα να αφορά ένα συγκεκριμένο θέμα. Στο ίδιο στάδιο χρησιμοποιούνται οι μέθοδοι TF-IDF και Gaussian. 4. Εξαγωγή της γεωγραφικής τοποθεσίας για το κάθε θέμα, κάνοντας χρήση εργαλείων για αναγνώριση επώνυμων οντοτήτων, POS ετικετών και επεξεργασία φυσικής γλώσσας, ώστε να προσδιοριστούν τα τοπονύμια. Επίσης χρησιμοποιούνται και τα metadata των αναρτήσεων. 5. Τα αποτελέσματα παρουσιάζονται στους χρήστες μέσω της διαδικτυακής εφαρμογής του συστήματος, με την οθόνη χωρισμένη σε δύο μέρη: Στο αριστερό μέρος εμφανίζονται οι θεματικές ομάδες με φθίνουσα σειρά προτεραιότητας και στο δεξί η γεωγραφική τους κατανομή σε χάρτη.
19	Open Domain Event Extraction from Twitter (Twical) [40]	1. Δεδομένης μιας ροής αναρτήσεων (raw stream of tweets), το προτεινόμενο σύστημα εντοπίζει και απομονώνει τις σημασιολογικές οντότητες (named entities) σε συσχέτιση με τις φράσεις που αφορούν γεγονότα και τη χρονική στιγμή δημοσίευσής τους. 2. Οι αναρτήσεις προσημαίνονται με ετικέτες (POS tags), εξάγονται οι οντότητες, αποσαφηνίζεται η χρονική τους σήμανση και τα γεγονότα που ανιχνεύονται χωρίζονται σε κατηγορίες. 3. Με το LinkLDA (Bayesian μοντέλο) γίνεται ομαδοποίηση των γεγονότων κάτω από ευρύτερες κατηγορίες 4. Τέλος υπολογίζεται η δύναμη της συσχέτισης μεταξύ κάθε οντότητας και του χρόνου δημοσίευσης με βάση τον αριθμό των αναρτήσεων στις οποίες εντοπίζεται, προκειμένου να αποσαφηνιστεί αν ένα γεγονός είναι «τάση» ή όχι. Προσδιορίζεται για κάθε γεγονός αν αυτό είναι σημαντικό, δηλαδή αν εμφανίζεται σε πολλές αναρτήσεις μέσα σε συγκεκριμένο χρονικό διάστημα.
20	An unsupervised framework of exploring events on twitter: filtering, extraction and categorization [41]	1. Στα πλαίσια της μεθοδολογίας που έχει σκοπό την ανίχνευση γεγονότων πρώτα κατασκευάζεται ένα λεξικό με τους κυριότερους όρους από τις ειδησεογραφικές ιστοσελίδες. 2. Η συλλογή δεδομένων γίνεται απευθείας από τη ροή αναρτήσεων του Twitter. 3. Στη συνέχεια γίνεται φιλτράρισμα των αναρτήσεων ανάλογα με το αν περιέχουν λέξεις μέσα από το συγκεκριμένο σύνολο και φιλτράρισμα βάσει χαρακτηριστικών (feature-based), με τη βοήθεια SVM μοντέλου. Από γλωσσολογική επεξεργασία γίνεται stemming, απόδοση ετικετών (POS) στις λέξεις, αναγνώριση επώνυμων οντοτήτων και η αντίστοιχη χαρτογράφησή τους (mapping).

		4. Για το κυρίως μέρος της ανίχνευσης (εξαγωγής) των γεγονότων και την ομαδοποίησή τους σε κατηγοριες χρησιμοποιείται ένα πιθανοτικό Bayesian μοντέλο, το LECM (Latent Event and Category Model), επέκταση του LEM. Γίνεται και ομαδοποίηση σε επίπεδο θέματος.
21	Real-time event detection for online behavioral analysis of big social data [42]	1. Η συλλογή δεδομένων γίνεται κάνοντας χρήση ενός όρου για αναζήτηση στο Twitter ώστε να περιορίζονται τα αποτελέσματα. 2. Τα δεδομένα κανονικοποιούνται σε ενιαία μορφή και τα περιττά στοιχεία αφαιρούνται. Στη συνέχεια εισάγονται σε βάση δεδομένων σε μορφή JSON format. 3. Τα κανονικοποιημένα δεδομένα αναλύονται για να εξαχθούν οι πιθανοί δημοφιλείς όροι. 4. Σημειώνεται η συχνότητα εμφάνισης των όρων και διάδοση της πληροφορίας για κάθε δείγμα. Τα στατιστικά στοιχεία του προηγούμενου βήματος χρησιμοποιούνται για να δημιουργήσουν σήματα. 5. Αφού έχουν εξαχθεί οι δημοφιλείς όροι και τα χαρακτηριστικά των αναρτήσεων, χτίζεται σημασιολογικό δίκτυο και γίνεται ομαδοποίηση η οποία καταλήγει σε δημοφιλή θέματα (events).
22	Microblog Topic Contagiousness Measurement and Emerging Outbreak Monitoring [37]	1. Λαμβάνεται από έτοιμο σύνολο δεδομένων ένα δείγμα αναρτήσεων για κάθε μήνα. 2. Με τη βοήθεια του Stanford Topic Modeling Toolbox γίνεται εφαρμογή PLDA μοντέλου θέματος για να ανιχνευθούν τα κεντρικά θέματα των αναρτήσεων. 3. Αποκλεισμός των αναρτήσεων οι οποίες δεν περιέχουν hashtags και εφαρμογή PLDA ξεχωριστά στα δύο σύνολα που προκύπτουν. Το σύστημα επιστρέφει πίνακες με την κατανομή των θεμάτων. 4. Υπολογισμός παραμέτρου διάδοσης R με χρήση του πακέτου ανοιχτού κώδικα Bayesian-inference Python package για κάθε θέμα - hashtag (SIR μοντέλο).

7. Βιβλιογραφία

[1] P. Kotler, *Marketing Management , Millenium Edition*, vol. 23, no. 6. 2000.

[2] R. E. Morgan and C. A. Strong, "Market orientation and dimensions of strategic orientation," *Eur. J. Mark.*, vol. 32, no. 11/12, pp. 1051–1073, 1998.

[3] ΤΖΩΡΤΖΑΚΗ Κ., *Αρχές Διοίκησης Μάρκετινγκ*. Αθήνα, 1993.

[4] A. Veisi, H. Rezvanfar, A. and Asadi, "Determining components of market orientation In aqua cultural higher education institutes," *Int. Rev. Public Non Profit Mark.*, vol. 4, pp. 81–89, 2007.

[5] Ι. Ν. Παρασκευόπουλος, *Μεθοδολογία επιστημονικής Έρευνας*. Αθήνα, 1993.

[6] Μ. Δ. Α. Σιώμκος Γεώργιος Ι., *Έρευνα Αγοράς*. Εκδόσεις Αθ.Σταμούλης, 2008.

[7] J. S. Brown, "Research that reinvents the corporation," *Harv. Bus. Rev.*, vol. 69, no. January-February, pp. 102–11, 1991.

[8] G. J. Avlonitis and S. P. Gounaris, "Marketing orientation and its determinants: an empirical analysis," *Eur. J. Mark.*, vol. 33, no. 11/12, pp. 1003–1037, 1999.

[9] S. Wright and J. L. Calof, "The quest for competitive, business and marketing intelligence," *Eur. J. Mark.*, vol. 40, no. 5/6, pp. 453–465, May 2006.

[10] E. Kyrkos, *Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων*. 2015.

[11] J. A. Balazs and J. D. Velásquez, "Opinion Mining and Information Fusion: A survey," *Inf. Fusion*, vol. 27, pp. 95–110, 2016.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[13] G. Miner, *Practical Text Mining and Statistical Analysis for Non-structured Text Data*. 2012.

[14] D. Laney, D. Management, and C. D. Volume, "META Delta," no. February 2001, 2005.

[15] KPMG International, "Going beyond the data: turning data from insights into value," pp. 1–32, 2015.

[16] G. Bello-orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, 2016.

[17] I. P. Cvijikj and F. Michahelles, "Monitoring trends on Facebook," *Proc. - IEEE 9th Int. Conf. Dependable, Auton. Secur. Comput. DASC 2011*, pp. 895–902, 2011.

[18] M. Injadat, F. Salo, and A. B. Nassif, "Data Mining Techniques in Social Media: A Survey Data Mining Techniques in Social Media: A Survey," *Neurocomputing*, vol. 214, no. I, pp. 1–17, 2016.

[19] A. Sapountzi and K. E. Psannis, "Social networking data analysis tools and challenges," *Futur. Gener. Comput. Syst.*, 2016.

[20] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," *2013 46th Hawaii Int. Conf. Syst. Sci.*, pp. 995–1004, 2013.

[21] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," *Sch. Comput. 9th. IT T Conf.*, p. 13, 2009.

[22] S. Brody and N. Diakopoulos, “Cooooooooooooooollllllllllll !!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs *** Preprint Version,” *Comput. Linguist.*, no. 2010, pp. 562–570, 2011.

[23] R. Dueñas-Fernández, J. D. Velásquez, G. L’Huillier, G. L. Huillier, and G. L’Huillier, “Detecting trends on the Web: A multidisciplinary approach,” *Inf. Fusion*, vol. 20, no. 1, pp. 129–135, 2014.

[24] J. C. L. Pinto, T. Chahed, E. Altman, J. Cesar, and L. Pinto, “Trend detection in social networks using Hawkes processes,” *Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM ’15*, pp. 1441–1448, 2015.

[25] A. Kontostathis, L. M. Galitsky, W. M. Pottenger, and D. J. Phelps, “A Survey of Emerging Trend Detection in Textual Data Mining,” pp. 1–44.

[26] A. Vakali, M. Giatsoglou, and S. Antaris, “Social networking trends and dynamics detection via a cloud-based framework design,” *Proc. 21st Int. Conf. companion World Wide Web - WWW ’12 Companion*, p. 1213, 2012.

[27] Y. M. Li and T. Y. Li, “Deriving market intelligence from microblogs,” *Decis. Support Syst.*, vol. 55, no. 1, pp. 206–217, 2013.

[28] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto, “Finding trendsetters in information networks,” *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’12*, pp. 1014–1022, 2012.

[29] Y. Han, B. Fang, and Y. Jia, “Predicting the topic influence trends in social media with multiple models,” *Neurocomputing*, vol. 144, pp. 463–470, 2014.

[30] Q. Dang, F. Gao, and Y. Zhou, “Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks,” *Expert Syst. Appl.*, vol. 57, pp. 285–295, 2016.

[31] M. Mathioudakis and N. Koudas, “Twittermonitor: trend detection over the twitter stream,” *SIGMOD ’10 Proc. 2010 ACM SIGMOD Int. Conf. Manag. data*, pp. 1155–1158, 2010.

[32] X. F. Yongtao Ye, Yajun Du, “Hot Topic Extraction Based on Chinese Microblog’s Features Topic Model,” *IEEE Int. Conf. Cloud Comput. Big Data Anal.*, pp. 348–353, 2016.

[33] S. Rill, D. Reinel, J. Scheidt, and R. V Zicari, “Knowledge-Based Systems PoliTwi : Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis,” *Knowledge-Based Syst.*, vol. 69, pp. 24–33, 2014.

[34] T. Zhu and J. Yu, “A Prerecognition Model for Hot Topic Discovery Based on Microblogging Data,” vol. 2014, 2014.

[35] H. Kim and S. Lee, “Discovering Hot Topics using Twitter Streaming Data Social Topic Detection and Geographic Clustering,” pp. 1215–1220, 2013.

[36] N. Glance, M. Hurst, and T. Tomokiyo, “BlogPulse: Automated trend discovery for weblogs,” *WWW 2004 Work. Weblogging Ecosyst. ACM New York*, vol. 2004, no. July, pp. 1–8, 2004.

[37] V. W. Chu, R. K. Wong, F. Chen, and C. Chi, “Microblog Topic Contagiousness Measurement and Emerging Outbreak Monitoring,” pp. 1099–1108, 2014.

[38] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, and J. Xia, “Event detection and popularity prediction in microblogging,” *Neurocomputing*, vol. 149, no. PC, pp. 1469–1480, 2015.

[39] J. Sankaranarayanan, B. E. Teitler, H. Samet, M. D. Lieberman, and J. Sperling, “TwitterStand : News in Tweets ∗ Categories and Subject Descriptors,” pp. 42–51, 2009.

[40] A. Ritter, O. Etzioni, and S. Clark, “Open Domain Event Extraction from Twitter,” pp. 1104–1112, 2012.

[41] D. Zhou and L. Chen, “An Unsupervised Framework of Exploring Events on Twitter : Filtering , Extraction and Categorization,” pp. 2468–2474.

[42] D. T. Nguyen and J. E. Jung, “Real-time event detection for online behavioral analysis of big social data,” *Futur. Gener. Comput. Syst.*, vol. 66, pp. 137–145, 2017.

[43] C. Paper, K. Ramachandran, A. Systems, and A. Chandra, “Sociopedia : An Interactive System for Event Detection and Trend Analysis for Twitter Data, June, 2015.

[44] E. Ferrara, P. De Meo, G. Fiumara, R. Baumgartner, P. De Meo, G. Fiumara, and R. Baumgartner, “Web data extraction, applications and techniques: A survey,” *Knowledge-Based Syst.*, vol. 70, pp. 301–323, 2014.

[45] B. Huang, Y. Yang, A. Mahmood, and H. Wang, “Microblog Topic Detection Based on LDA Model and Single-Pass Clustering,” pp. 166–171, 2012.

[46] D. H. Chau, S. Pandit, S. Wang, and C. Faloutsos, “Parallel crawling for online social networks,” in *Proceedings of the 16th international conference on World Wide Web - WWW ’07*, 2007, p. 1283.

[47] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, “Crawling Facebook for Social Network Analysis Purposes,” pp. 0–7, 2011.

[48] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Walking in facebook: A case study of unbiased sampling of OSNs,” in *Proceedings - IEEE INFOCOM*, 2010.

[49] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, “User Interactions in Social Networks and Their Implications,” *Proc. 4th {ACM} Eur. Conf. Comput. Syst.*, pp. 205–218, 2009.

[50] S. Ye, J. Lang, and F. Wu, “Crawling online social graphs,” in *Advances in Web Technologies and Applications - Proceedings of the 12th Asia-Pacific Web Conference, APWeb 2010*, 2010, pp. 236–242.

[51] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter , a Social Network or a News Media?,” *Int. World Wide Web Conf. Comm.*, pp. 1–10, 2010.

[52] K. Toutanova and C. D. Manning, “Enriching the knowledge sources used in a maximum entropy part-of-speech tagger,” *Proc. 2000 Jt. SIGDAT Conf. Empir. methods Nat. Lang. Process. very large corpora held conjunction with 38th Annu. Meet. Assoc. Comput. Linguist. -*, vol. 13, pp. 63–70, 2000.

[53] Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng, “Hashtag Graph Based Topic Model for Tweet Mining,” *2014 IEEE Int. Conf. Data Min.*, pp. 1025–1030, 2014.

[54] X. Cheng, X. Yan, Y. Lan, and J. Guo, “BTM: Topic modeling over short texts,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2928–2941, 2014.

[55] S. Deerwester, S. T. Dumais, and R. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[56] T. Hofmann, “Probabilistic latent semantic indexing,” *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 50–57, 1999.

[57] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[58] D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models,” *IEEE Signal Process. Mag.*,

vol. 27, no. 6, pp. 55–65, 2010.

[59] S. Gerard and M. J. Michael, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, NY, USA, 1986.

[60] A. C. M. Transactions and I. Technology, "A Graph Analytical Approach for Topic Detection," no. December 2013, 2016.

[61] H. Wang and Y. Ohsawa, "IdeaGraph: Turning Data into Human Insights for Collective Intelligence," in *Advances in Intelligent Systems and Computing*, 2013.

[62]] W. Medhat et Al., "Sentiment analysis algorithms and applications: a survey," *Ain Shams Eng. J*, 2014.

[63] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Syst.*, vol. 89, pp. 14–46, 2015.

[64] A. Balahur, "Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types," University of Alicante, Spain, 2011.

[65] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, pp. 1–8, 2012.

[66] M. Adedoyin-olowe, M. M. Gaber, and F. Stahl, "A Survey of Data Mining Techniques for Social Network Analysis," *Int. J. Res. Comput. Eng. Electron.*, vol. 3, no. 6, pp. 1–8, 2014.