ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΚΑΙ ΔΙΑΔΙΚΤΥΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Κ
ΕΦΑΡΜΟΓΕΣ

«ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΠΤΥΞΗ ΤΗΛΕΜΑΤΙΚΩΝ ΕΦΑΡΜΟΓΩΝ ΣΕ ΕΞΥΠΝΕΣ ΠΟΛΕΙΣ»

Μεταπτυχιακή Διπλωματική Εργασία

Ρομποτή Ευσταθία

Α.Μ: 12408

Αθήνα, 2018

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΚΑΙ ΔΙΑΔΙΚΤΥΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Κ
ΕΦΑΡΜΟΓΕΣ

Τριμελής Εξεταστική Επιτροπή

Όνομα Πρώτου Καθηγητή (Επιβλέπων)

Γιώργος Δημητρακόπουλος

Επίκουρος καθηγητής, Χαροκόπειο πανεπιστήμιο

Όνομα Δεύτερου Καθηγητή

Γιώργος Μπράβος

Λέκτορας, Χαροκόπειο πανεπιστήμιο

Όνομα Τρίτου Καθηγητή

Κωνσταντίνος Τσερπές

Επίκουρος καθηγητής, Χαροκόπειο πανεπιστήμιο

Η Ρομποτή Ευσταθία δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

Acknowledgements

This diploma thesis has been developed in the framework of the Postgraduate Program "Informatics and Telematics" in the direction of "MSc in Web Engineering", at Department of Informatics and Telematics of Harokopio University, under the supervision of George Dimitrakopoulos, George Bravos and Konstantinos Tserpes.

I am very thankful for the opportunity to work with them, for their fruitful and comprehensive discussions, for the scientific knowledge that they shared with me throughout this master thesis.

I would like to thank also to the departmental staff for their help and encouragement.

Contents

Περίληψη

Οι έξυπνες πόλεις βασίζονται σε μετρήσεις σε πραγματικό χρόνο και δεδομένα που συλλέγονται από μεγάλο αριθμό ετερογενών φυσικών αισθητήρων που αναπτύσσονται σε όλη την πόλη. Ωστόσο, οι φυσικοί αισθητήρες είναι γεμάτοι με προκλήσεις διαλειτουργικότητας, αξιοπιστίας και διαχείρισης. Επιπλέον, αυτοί οι αισθητήρες δεν είναι σε θέση να αντιληφθούν τις απόψεις και τις συναισθηματικές αντιδράσεις των πολιτών που επηρεάζουν αμετάβλητα τις πρωτοβουλίες έξυπνων πόλεων. Καθημερινά, εκατομμύρια κάτοικοι και επισκέπτες μιας πόλης μοιράζονται τις παρατηρήσεις, τις σκέψεις, τα συναισθήματα και τις εμπειρίες τους μέσω των ενημερώσεων των κοινωνικών μέσων ενημέρωσης. Με την άνοδο των κοινωνικών μέσων ενημέρωσης, οι άνθρωποι αποκτούν και μοιράζονται πληροφορίες σχεδόν άμεσα σε 24ωρη βάση. Πολλοί ερευνητικοί τομείς προσπάθησαν να αποκτήσουν πολύτιμες πληροφορίες από αυτούς τους μεγάλους όγκους διαθέσιμων δεδομένων που δημιουργούν οι χρήστες. Οι "ανθρώπινοι αισθητήρες", δηλαδή οι πολίτες μοιράζονται πληροφορίες για το περιβάλλον τους μέσω των κοινωνικών μέσων ενημέρωσης, μπορούν να συμπληρώνουν, να συμπληρώνουν τις πληροφορίες που μετρούνται με φυσικούς αισθητήρες. Τα κοινωνικά μέσα μπορούν να παρέχουν έξυπνη νοημοσύνη. Αυτή η έξυπνη ευφυΐα επιτρέπει στην έξυπνη πόλη να βελτιώσει την ποιότητα ζωής και να διαχειριστεί τη λειτουργία της εφοδιαστικής με έξυπνο τρόπο.

Η έννοια των κοινωνικών μέσων μαζικής ενημέρωσης βρίσκεται στην κορυφή της ατζέντας για πολλά στελέχη επιχειρήσεων σήμερα. Οι υπεύθυνοι λήψης αποφάσεων, καθώς και οι σύμβουλοι, προσπαθούν να εντοπίσουν τρόπους με τους οποίους οι επιχειρήσεις μπορούν να κάνουν αποδοτική χρήση εφαρμογών όπως το Wikipedia, το YouTube, το Facebook, το Second Life και το Twitter. Κάτω από την ορολογία των κοινωνικών μέσων μαζικής ενημέρωσης, θέτουμε όλους τους ιστότοπους / δίκτυα όπου μπορούμε να διαβάζουμε και να γράφουμε "περιεχόμενο που δημιουργείται από το χρήστη", ενώ οι σημαντικές αναφορές των καταναλωτών (OCR) αποτελούν σημαντικό μέρος αυτού του περιεχομένου.

Παρόλο που οι OCR βοήθησαν τους καταναλωτές να μάθουν για τα πλεονεκτήματα και τις αδυναμίες των διαφόρων προϊόντων και να βρουν εκείνα που ταιριάζουν καλύτερα στις ανάγκες τους, εισάγουν μια πρόκληση για τις επιχειρήσεις να τα αναλύουν εξαιτίας του όγκου, της ποικιλίας, της ταχύτητας και της ακρίβειας.

Οι προγνωστικοί παράγοντες της αναγνωσιμότητας και της χρησιμότητας του OCR χρησιμοποιώντας μια προσέγγιση εξόρυξης για μεγάλες αναλύσεις δεδομένων έχουν ερευνηθεί μόνο τα τελευταία χρόνια. Τα ευρήματα έως τώρα δείχνουν αποτελέσματα που θα μπορούσαν να χρησιμοποιηθούν ως βάση για νέα επιχειρηματικά μοντέλα και προσαρμογές σε διάφορες διαδικασίες που ακολουθούνται από μεγάλες επιχειρήσεις, συμπεριλαμβανομένων των διαδικασιών και των μοντέλων που χρησιμοποιούνται στον τομέα της έξυπνης εφοδιαστικής. Παρ 'όλα αυτά, οι τρέχουσες μέθοδοι που χρησιμοποιούνται για τη διαλογή OCR μπορεί να προκαλέσουν τόσο την αναγνωσιμότητα όσο και την εξυπηρετικότητα τους. Είναι απαραίτητο να αναπτυχθούν κλιμακούμενα αυτοματοποιημένα συστήματα ταξινόμησης και ταξινόμησης μεγάλων δεδομένων OCR τα οποία θα ωφελήσουν τόσο τους πωλητές όσο και τους καταναλωτές.

Με στόχο την εξαγωγή γνώσεων από ροές κοινωνικών μέσων που θα μπορούσαν να είναι χρήσιμες στο πλαίσιο έξυπνων συστημάτων μεταφορών και έξυπνων πόλεων, σχεδιάσαμε και αναπτύξαμε ένα εργαλείο το οποίο θα είναι σε θέση να ικανοποιήσει τα εξής:

- Να κατανοήσει της συμπεριφοράς των καταναλωτών σχετικά με τη διαδικασία εφοδιαστικής

- Να κατανοήσει τα προβλήματα των τμημάτων της ηλεκτρονικής διαδικασίας αγοράς

- Να προσδιορίσει τα κύρια προβλήματα που εντοπίζονται από τους πελάτες

- Να δώσει στις επιχειρήσεις ηλεκτρονικού εμπορίου την ευκαιρία να βελτιώσουν τμήματα της διαδικασίας εφοδιαστικής, καθώς και να αναπτύξουν νέα επιχειρηματικά μοντέλα που θα είναι σε θέση να ελαχιστοποιούν το κόστος μεταφοράς / υλικοτεχνικής υποστήριξης.

Στο πρώτο κεφάλαιο παρουσιάζουμε τα χαρακτηριστικά και τις τεχνολογίες IoT και Smart Cities καθώς το IoT γίνεται μέρος μιας έξυπνης πόλης. Στο δεύτερο κεφάλαιο παρουσιάζουμε το Logistics and Intelligent Transportation System (ITS), καθώς οι έξυπνες πόλεις χρειάζονται έξυπνες μεταφορικές υπηρεσίες. Στο τρίτο κεφάλαιο κάνουμε μια εισαγωγή στο ηλεκτρονικό εμπόριο. Εξηγούμε τη σημασία της εφοδιαστικής και των μεγάλων δεδομένων για το ηλεκτρονικό εμπόριο. Καταλήγουμε ότι σε ορισμένες ειδικές απαιτήσεις που αποκάλυψαν πολυάριθμους τομείς ότι η πιθανή μεγάλη συμβολή και εφαρμογή δεδομένων θα μπορούσε να επιφέρει σημαντικές βελτιώσεις όσον αφορά το κόστος και την εξυπηρέτηση των πελατών. Στο κεφάλαιο 4 παρουσιάζουμε γενικές τεχνικές και αλγόριθμους εξόρυξης κειμένων, ταξινόμησης κειμένου και ομαδοποίησης κειμένων, όπως στο Κεφάλαιο 5, με βάση τις τεχνικές εξόρυξης κειμένου και σημασίας, προσδιορίζουμε τα προβλήματα των χρηστών που σχετίζονται με τη διαδικασία εφοδιαστικής, μέσω της ανάλυσης των δεδομένων των κοινωνικών μέσων.

Λέξεις κλειδιά: Έξυπνες πόλεις, Ηλεκτρονικό εμπόριο, εξόρυξη κειμένου.

Introduction

Smart city initiatives rely on real-time measurements and data collected by a large number of heterogenous physical sensors deployed throughout a city. Physical sensors, however, are fraught with interoperability, dependability and management challenges. Furthermore, these sensors are unable to sense the opinions and emotional reactions of citizens that invariably impact smart city initiatives. Yet every day, millions of dwellers and visitors of a city share their observations, thoughts, feelings and experiences through social media updates. With the rise of Social Media, people obtain and share information almost instantly on a 24/7 basis. Many research areas have tried to gain valuable insights from these large volumes of freely available user generated content. “Human sensors”, namely, citizens share information about their surroundings via social media can supplement, complement, or even replace the information measured by physical sensors. Social media can give predictive intelligence. This predictive intelligence allows the Smart City to improve quality of life and managing logistics operation with smart way.

The concept of social media is top of the agenda for many business executives today. Decision makers, as well as consultants, try to identify ways in which firms can make profitable use of applications such as Wikipedia, YouTube, Facebook, Second Life, and Twitter. Under the umbrella of the term social media, we put all the sites/networks in which we may read and write “user generated content”, with online consumer reviews (OCRs) being a significant part of that content.

Although OCRs have helped consumers to know about the strengths and weaknesses of different products and find the ones that best suit their needs, they introduce a challenge for businesses to analyse them because of their volume, variety, velocity and veracity. The predictors of readership and helpfulness of OCR using a sentiment mining approach for big data analytics has been investigated only in the past few years; findings up to now indicate results that could be used as the basis for new business models and adjustments in several procedures followed by large enterprises, including the processes and models used in the smart logistics field. Nevertheless, current methods used for sorting OCR may bias both their readership and helpfulness; thus, it is necessary to develop scalable automated systems for sorting and classification of big OCR data which will benefit both vendors and consumers.

With the goal of extracting knowledge from social media streams that might be useful in the context of intelligent transportation systems and smart cities, we designed and developed a tool which will be able to fulfil the following:

- Understand consumers’ behavior relevant to logistics process
- Understand the problematic parts of the online purchasing procedure

- Identify the main problems identified by customers
- Give to e-commerce firms the opportunity to ameliorate parts of its logistics procedure, as well as to develop new business models which will be able to minimize their “transportation”/ logistics cost.

In the first Chapter we present IoT and Smart Cities features and technologies as IoT becomes part of a Smart City. In the second Chapter we present Logistics and Intelligent Transportation System (ITS) as Smart Cities need smart transport services. In Chapter three we make an introduction to e-commerce. We explain the importance of logistics and big data on e-commerce. We conclude into some specific requirements that revealed numerous areas that potential big data contribution and implementation could bring significant improvements in terms of cost and customer service. In chapter four we present general techniques and algorithms of text mining, text classification and text clustering as in Chapter 5 based on text mining and meaning techniques we identify users’ problems relevant to the logistics procedure, via the analysis of social media data.

Key words: Smart Cities, Transportation, e-commerce, Text mining, Meaning mining.

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1 Smart Cities backbone: The Internet of Things

The Internet of Things[1] represents a vision in which the Internet extends into the real world embracing everyday objects. Physical items are no longer disconnected from the virtual world but can be controlled remotely and can act as physical access points to Internet services.

IoT (Internet of Things) is an advanced automation and analytics system which exploits networking, sensing, big data, and artificial intelligence technology to deliver complete systems for a product or service. These systems allow greater transparency, control, and performance when applied to any industry or system.

IoT systems have applications across industries through their unique flexibility and ability to be suitable in any environment. They enhance data collection, automation, operations, and much more through smart devices and powerful enabling technology.

1.1.1 IoT – Overview

IoT systems allow users to achieve deeper automation, analysis, and integration within a system. They improve the reach of these areas and their accuracy. IoT utilizes existing and emerging technology for sensing, networking, and robotics. IoT exploits recent advances in software, falling hardware prices, and modern attitudes towards technology. Its new and advanced elements bring major changes in the delivery of products, goods, and services; and the social, economic, and political impact of those changes.

1.1.2 IoT – Key Features

From a technical point of view, the Internet of Things is not the result of a single novel technology; instead, several complementary technical developments provide capabilities that taken together help to bridge the gap between the virtual and physical world. A brief review of these features is given below:

AI: IoT essentially makes virtually anything “smart”, meaning it enhances every aspect of life with the power of data collection, artificial intelligence algorithms, and networks. This can mean something as simple as enhancing your refrigerator and cabinets to detect when milk and your favorite cereal run low, and to then place an order with your preferred grocer.

Addressability: Within an Internet of Things, objects can be located and addressed via discovery, look-up or name services, and hence remotely interrogated or configured.

Connectivity: New enabling technologies for networking, and specifically IoT networking, mean networks are no longer exclusively tied to major providers. Networks can exist on a much smaller and cheaper scale while still being practical. IoT creates these small networks between its system devices.

Communication and cooperation: Objects have the ability to network with Internet resources or even with each other, to make use of data and services and update their state. Wireless technologies such as GSM and UMTS, Wi-Fi, Bluetooth, ZigBee and various other wireless networking standards currently under development, particularly those relating to Wireless Personal Area Networks (WPANs), are of primary relevance here.

Identification: Objects are uniquely identifiable. RFID, NFC (Near Field Communication) and optically readable bar codes are examples of technologies with which even passive objects which do not have built-in energy resources can be identified (with the aid of a “mediator” such as an RFID reader or mobile phone). Identification enables objects to be linked to information associated with the particular object and that can be retrieved from a server, provided the mediator is connected to the network

Sensors: IoT loses its distinction without sensors. They act as defining instruments which transform IoT from a standard passive network of devices into an active system capable of real-world integration. Objects collect information about their surroundings with sensors, record it, forward it or react directly to it.

Actuation: Objects contain actuators to manipulate their environment (for example by converting electrical signals into mechanical movement). Such actuators can be used to remotely control real-world processes via the Internet.

Embedded information processing: Smart objects feature a processor or microcontroller, plus storage capacity. These resources can be used, for example, to process and interpret sensor information, or to give products a “memory” of how they have been used.

Small Devices: Devices, as predicted, have become smaller, cheaper, and more powerful over time. IoT exploits purpose-built small devices to deliver its precision, scalability, and versatility.

User interfaces: Smart objects can communicate with people in an appropriate manner (either directly or indirectly, for example via a smartphone). Innovative interaction paradigms are relevant here, such as tangible user interfaces, flexible polymer-based displays and voice, image or gesture recognition methods.

Localization: Smart things are aware of their physical location, or can be located. GPS or the mobile phone network are suitable technologies to achieve this, as well as ultrasound time measurements, UWB (Ultra-Wide Band), radio beacons (e.g. neighboring WLAN base stations or RFID readers with known coordinates) and optical technologies.

IoT – Hardware

The hardware[1] utilized in IoT systems includes devices for a remote dashboard, devices for control, servers, a routing or bridge device, and sensors. These devices manage key tasks and functions such as system activation, action specifications, security, communication, and detection to support-specific goals and actions.

IoT – Sensors

The most important hardware in IoT might be its sensors. These devices consist of energy modules, power management modules, RF modules, and sensing modules. RF modules manage communications through their signal processing, WiFi, ZigBee, Bluetooth, radio transceiver, duplexer, and BAW. The sensing module manages sensing through assorted active and passive measurement devices. Here is a list of some of the measurement devices used in IoT:

Devices	
Accelerometers	Temperature sensors
Magnetometers	Proximity sensors
Gyroscopes	Image sensors
Acoustic sensors	Light sensors
Pressure sensors	Gas RFID sensors
Humidity sensors	Micro flow sensors

Standard Devices

The desktop, tablet, and cellphone remain integral parts of IoT as the command center and remotes.

The **desktop** provides the user with the highest level of control over the system and its settings.

The **tablet** provides access to the key features of the system in a way resembling the desktop, and also acts as a remote.

The **cellphone** allows some essential settings modification and also provides remote functionality. Other key connected devices include standard network devices like **routers** and **switches**.

IoT – Software

IoT software addresses its key areas of networking and action through platforms, embedded systems, partner systems, and middleware. These individual and master applications are responsible for data collection, device integration, real-time analytics, and application and process extension within the IoT network. They exploit integration with critical business systems (e.g., ordering systems, robotics, scheduling, and more) in the execution of related tasks.

Data Collection

This software manages sensing, measurements, light data filtering, light data security, and aggregation of data. It uses certain protocols to aid sensors in connecting with real-time, machine-to-machine networks. Then it collects data from multiple devices and distributes it in accordance with settings. It also works in reverse by distributing data over devices. The system eventually transmits all collected data to a central server.

Device Integration

Software supporting integration binds (dependent relationships) all system devices to create the body of the IoT system. It ensures the necessary cooperation and stable networking between devices. These applications are the defining software technology of the IoT network because without them, it is not an IoT system. They manage the various applications, protocols, and limitations of each device to allow communication.

Real-Time Analytics

These applications take data or input from various devices and convert it into viable actions or clear patterns for human analysis. They analyze information based on various settings and designs in order to perform automation-related tasks or provide the data required by industry.

Application and Process Extension

These applications extend the reach of existing systems and software to allow a wider, more effective system. They integrate predefined devices for specific purposes such as allowing certain mobile devices or engineering instruments access. It supports improved productivity and more accurate data collection.

IoT − Technology and Protocols

IoT primarily exploits standard protocols and networking technologies. However, the major enabling technologies and protocols of IoT are RFID, NFC, low-energy Bluetooth, low-energy wireless, low-energy radio protocols, LTE-A, and WiFi-Direct. These technologies support the specific networking functionality needed in an IoT system in contrast to a standard uniform network of common systems.

NFC and RFID

RFID (radio-frequency identification) and NFC (near-field communication) provide simple, low energy, and versatile options for identity and access tokens, connection bootstrapping, and payments.

RFID technology employs 2-way radio transmitter-receivers to identify and track tags associated with objects.

NFC consists of communication protocols for electronic devices, typically a mobile device and a standard device.

Low-Energy Bluetooth

This technology supports the low-power, long-use need of IoT function while exploiting a standard technology with native support across systems.

Low-Energy Wireless

This technology replaces the most power hungry aspect of an IoT system. Though sensors and other elements can power down over long periods, communication links (i.e., wireless) must remain in listening mode. Low-energy wireless not only reduces consumption, but also extends the life of the device through less use.

Radio Protocols

ZigBee, Z-Wave, and Thread are radio protocols for creating low-rate private area networks. These technologies are low-power, but offer high throughput unlike many similar options. This increases the power of small local device networks without the typical costs.

LTE-A

LTE-A, or LTE Advanced, delivers an important upgrade to LTE technology by increasing not only its coverage, but also reducing its latency and raising its throughput. It gives IoT a tremendous power through expanding its range, with its most significant applications being vehicle, UAV, and similar communication.

WiFi-Direct

WiFi-Direct eliminates the need for an access point. It allows P2P (peer-to-peer) connections with the speed of WiFi, but with lower latency. WiFi-Direct eliminates an element of a network that often bogs it down, and it does not compromise on speed or throughput.

1.1.3 IoT – Common Uses

IoT has applications across all industries and markets. It spans user groups from those who want to reduce energy use in their home to large organizations who want to streamline their operations. It proves not just useful, but nearly critical in many industries as technology advances and we move towards the advanced automation imagined in the distant future.

IoT – Transportation Applications

At every layer of transportation, IoT provides improved communication, control, and data distribution. These applications include personal vehicles, commercial vehicles, trains, UAVs, and other equipment. It extends throughout the entire system of all transportation elements such as traffic control, parking, fuel consumption, and more.

Rails and Mass Transit

Current systems deliver sophisticated integration and performance, however, they employ older technology and approaches to MRT. The improvements brought by IoT deliver more complete control and monitoring. This results in better management of overall performance, maintenance issues, maintenance, and improvements. Mass transit options beyond standard MRT suffer from a lack of the integration necessary to transform them from an option to a dedicated service. IoT provides an inexpensive and advanced way to optimize performance and bring qualities of MRT to other transportation options like buses. This improves services and service delivery in the areas of scheduling, optimizing transport times, reliability, managing equipment issues, and responding to customer needs.

Road

The primary concerns of traffic are managing congestion, reducing accidents, and parking. IoT allows us to better observe and analyze the flow of traffic through devices at all traffic observation points. It aids in parking by making storage flow transparent when current methods offer little if any data.

Accidents typically result from a number of factors, however, traffic management impacts their frequency. Construction sites, poor rerouting, and a lack of information about traffic status are all issues that lead to incidents. IoT provides solutions in the form of better information sharing with the public, and between various parties directly affecting road traffic.

Automobile

Many in the automotive industry envision a future for cars in which IoT technology makes cars "smart," attractive options equal to MRT. IoT offers few significant improvements to personal vehicles. Most benefits come from better control over related infrastructure and the inherent flaws in automobile transport; however, IoT does improve personal vehicles as personal spaces. IoT brings the same improvements and customization to a vehicle as those in the home.

Commercial Transportation

Transportation benefits extend to business and manufacturing by optimizing the transport arm of organizations. It reduces and eliminates problems related to poor fleet management through better analytics and control such as monitoring idling, fuel consumption, travel conditions, and travel time between points. This results in product transportation operating more like an aligned service and less like a collection of contracted services.

IoT − Media, Marketing, & Advertising

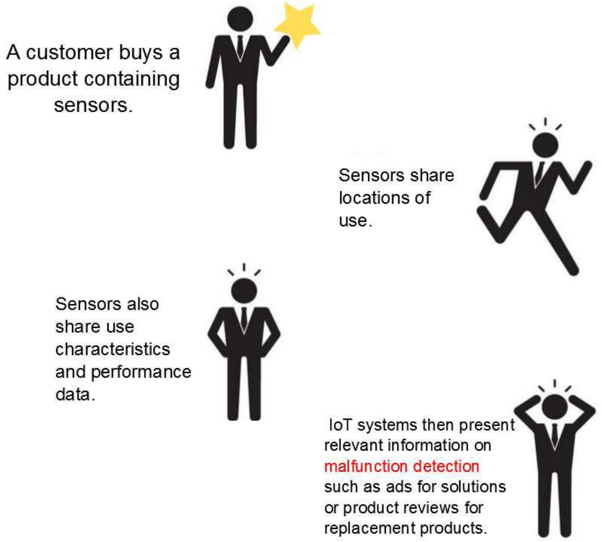
The applications of IoT in media and advertising involve a customized experience in which the system analyzes and responds to the needs and interests of each customer. This includes their general behavior patterns, buying habits, preferences, culture, and other characteristics.

Marketing and Content Delivery

IoT functions in a similar and deeper way to current technology, analytics, and big data. Existing technology collects specific data to produce related metrics and patterns over time, however, that data often lacks depth and accuracy. IoT improves this by observing more behaviors and analyzing them differently.

- This leads to more information and detail, which delivers more reliable metrics and
  patterns.

- It allows organizations to better analyze and respond to customer needs or preferences.
- It improves business productivity and strategy and improves the consumer experience by only delivering relevant content and solutions.


A customer buys a product containing sensors.
Sensors share locations of use.
Sensors also share use characteristics and performance data.
IoT systems then present relevant information on malfunction detection such as ads for solutions or product reviews for replacement products.

1.2 Smart Cities

Defining the Terms

There is no common consensus about what “smart” really means in the context of the information and communications technology (ICT)[1]. Although this term has become fashionable, it is also broadly used as a synonym of almost anything considered to be modern and intelligent. Smart, in purely definitional terms, has many synonyms, including percipient, astute, shrewd, and quick[1]. Moreover, smart is synonymous to efficient, when it is linked to devices[2]. Moreover, the term smart refers to ideas and people that provide clever insights but it has been adopted more recently in city planning through the cliché smart growth[2]. Growth can be seen as city sprawl, population increase or local economic upgrade, while smart growth implies the achievement of greater city efficiency through coordinating the forces that lead to growth: transportation, land speculation, conservation, and economic development[6].

Similarly, it is not easy to locate a common definition for the term city, while most people can conceptualize it according to individual experiences. A city is considered as an urban area, which according the United Nations (2005) typically begins with a population density of 1500 people per square mile but it varies across countries. Cities range according to their agglomeration from localities or villages (e.g., Greenland and Iceland) of 200–1000 inhabitants; to communities (or communes) of 1000–2500 people (e.g., Africa), to towns or places (e.g., Canada) or cantons with more than 400 (e.g., Albania) and less than 10,000 inhabitants (e.g., Greece); to cities with a population over 10,000 and 1.5 million inhabitants; and megacities with a population that exceeds 1.5 million people. Some cities are also called global or international due to their impact that attracts inhabitants beyond the country or even from all over the world. Small and medium-sized cities compete for resources against larger and better-equipped ones, while they all have peers (e.g., cities with similar characteristics)[1]. Another indicative definition says that "city is an urban community falling under a specific administrative boundary"[1,9], which shows that a city needs some model of governance. Community is a group of people with an arrangement of responsibilities, activities and relationships"[10]. Moreover, "a city is a system of systems with a unique history and set in a specific environmental and societal context. In order for it to flourish, all the key city actors need to work together, utilizing all of their resources, to overcome the challenges and grasp the opportunities that the city faces"[9].

1.2.1 What Is Smart City?

It would be normal for someone to consider that smart city comes up from the combination of the above definitions: an urban space that is surrounded by or is embedded with "smart systems" or a city with ideas and people that provide clever insights. Smart systems should not be limited to ICT-based ones, but intelligence can refer even to creative design or new organizations etc. In this regard, the "smartness" of a city describes its ability to bring together all its resources, to effectively and seamlessly achieve the goals and fulfil the purposes it has set itself[9]. However, if someone seeks for a clear definition for smart city, he will fail to locate one and instead, he will retrieve many alternatives, which generate an ambiguous meaning.

The European Commission programs FP7-ICT and CIP ICT-PSP approaches smart city as a "user-driven open innovation environment"[11], where city is seen as a platform that enhances citizen engagement and their willing to "co-create". "Openness" is being conceptualized in terms to apply various forms of relationships between people, services, infrastructure and technology[12]. Open public services facilitate the coordination of people's participatory "living-playing-working" activities, while open-service oriented business models work according to open industry standards (in terms of infrastructure and technology)[12]. In this respect, open innovation systems promote high quality social interactions (e.g., within communities), which enhance citizen engagement and participatory decision making. Finally, it is important to mention how standardization bodies—at least the international ones—define the smart city: the International Telecommunications Union (ITU)[13,14] emphasizes on ICT and considers a smart sustainable city as an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social and environmental aspects. Similarly, the International Standards Organization (ISO)[14] recognizes smart city as a new concept and a new model, which applies the new generation of information technologies, such as the internet of things, cloud computing, big data and space/geographical information integration, to facilitate the planning, construction, management and smart services of cities. Moreover, it defines smart city objective to pursue: convenience of the public services; delicacy of city management; livability of living environment; smartness of infrastructures; long-term effectiveness of network security. Furthermore, the British Standards[15] concerns the smart city as the effective integration of physical, digital and human systems in the built environment to deliver a sustainable, prosperous and inclusive future for its citizens.

1.2.2 Smart City Ecosystem & Classification

Some of the smart city challenges have been already identified: providing an economic base; building efficient urban infrastructure; improving the quality of life and place; ensuring social integration; conserving natural environmental qualities, and; guaranteeing good governance[16]. Moreover, these definitions demonstrate that scholars conceptualize smart city with alternative approaches. In this respect[17], performed a comparative analysis on existing smart city conceptual models. These models synthesize a smart city ecosystem, which consists of eight (8) components that establish a cyber-physical integration and—with the incorporation of standardization perspectives—concern:

1. Smart infrastructure: city facilities (e.g., water and energy networks, streets, buildings etc.) with embedded smart technology (e.g., sensors, smart grids etc.).

2. Smart Transportation (or smart mobility): transportation networks with enhanced embedded real time monitoring and control systems.

3. Smart Environment: innovation and ICT incorporation for natural resource protection and management (waste management systems, emission control, recycling, sensors for pollution monitoring etc.).

4. Smart Services: utilization of technology and ICT for health, education, tourism, safety, response control (surveillance) etc. service provision across the entire city.

5. Smart Governance: smart government establishment in the urban space, accompanied by technology for service delivery, participation and engagement.

6. Smart People: measures that enhance people creativity and open innovation.

7. Smart Living: innovation for enhancing quality of life and livability in the urban space.

8. Smart Economy: technology and innovation for strengthening business development, employment and urban growth. These components are interconnected and require data collection and ICT infrastructure, to be embedded within city hard infrastructure to deliver smart services to city actors, while governance is necessary in order for the subsystems to be orchestrated and succeed in smart city mission.

The extreme smart city growth that has been performed during the last 20 years has created various alternative smart city types. For a beginning[18], classify smart cities in market-driven groups: “GreenFields” and “Brownfields” that display the size (large-scale cases compared to small-scale ones) of the smart city project; and to four different “box” types according to project organization and business model:

• Information Technology (IT) box: a private company initiates the smart city and private funding business model;

• Dream box: public-private partnership (PPP) for project definition and respective business model;

• Fragmented box: many projects initiated by various stakeholders with little or no integration; and

• Black box: initiated and managed by (local, state or national) Governments or public agencies, with “invited” companies to enter this ecosystem. Additionally[19], made an analysis of 34 different smart cities and discovered alternatives that vary with regard to the ICT that has been embedded within the city and defines an alternative smart adjective to city. These alternatives determine several smart city classes, which range and mainly address the adjective that describes the particular ICT that is installed in the city. More specifically, the following classes can be located[19]. Web or Virtual Cities: offer local information, online chatting and meeting rooms, and city augmented reality navigation via the Web. Some indicative cases concern: America-On-Line (AOL) Cities (1997), Kyoto, Japan (1998– 2001), Bristol, U.S.A. (1997) and Amsterdam (1997).

• Knowledge Bases[20] or Knowledge Cities: are digital public repositories with crowd sourcing options accessible via the Internet and via text-TV (Copenhagen Base (1989); CraigmillarCommunity Information Service, Scotland (1994); and Blacksburg Knowledge Democracy). Later approaches[21,22], define knowledge city as locally focused innovation, science and creativity within the context of an expanding knowledge economy and society. This later approach has been followed by Melbourne.

• Broadband City/Broadband Metropolis: describe fiber optic backbones in the urban area, which enable the interconnection of households and of local enterprises to ultra-high speed networks. Seoul, S. Korea (1997); Beijing, China (1999); Helsinki (1995); Geneva-MAN, Switzerland (1994) (van Bastelaer 1998); and Antwerp comprised this category. Mobile/Wireless/Ambient Cities are wireless broadband networks accessible across the city or in some districts. New York City (1994); Kista Science City/Stockholm (2002) and Florence, Italy (2006) were the identified representative members.

- The Digital or Information City describes an ICT environment across the city that is built to deal with: (a) local needs and transactions, (b) the transformation of the local community to a local information society, and (c) sustainable local development. Hull, U.K. (2000); Cape Town, South Africa (2000); Trikala, Greece (2003); Tampere, Finland (2003); Knowledge Based Cities, Portugal (1995); and Austin, U.S.A. (1995—today) are members of this group.

- The Ubiquitous City extends the digital or information city in enabling ubiquitous service provision and data flow from anywhere to everyone. New Sondgo, S. Korea (2008); Dongtan, S. Korea (2005); Osaka, Japan (2008); Manhattan Harbour, Kentucky, U.S.A. (2010); Masdar, United Arab Emirates (2008); and Helsinki Arabianranta, Finland (2005) are some representatives.

- The smart city came to extend ubiquitous city in a sense that emphasized social infrastructure (human and social capital, named the dimension of people) of the city (Lee et al. 2014). This approach offers broadband and media infrastructures for business growth too. Taipei, Taiwan (2004); Tianjin, China (2007); Barcelona, Spain (2000); Brisbane, Australia (2004); Malta (2007); Kochi, India (2007); and Dubai (1999—today) were labeled “smart” from their initial appearance.

- Finally, the Eco City extends ubiquitous city with a service agenda that respects the physical landscape of the city or in other words it capitalizes the ICT for sustainable growth and for environmental protection.

1.2.3 Smart Services

An analysis[23] of the types of smart services that 29 of the above cases offer and structured nine (9) smart service groups (SG), inspired by market-driven[18] groups:

- SG1: e-Government services (City Administration market-driven group) concern typical public transactions (offered by digital, smart and ubiquitous city classes).

- SG2: e-democracy services (City Administration market-driven group), like dialogues, consultation, polling and voting than enhance citizen engagement (offered by virtual, digital, smart and ubiquitous city classes).

- SG3: smart business services (Real estate market-driven group), that concern business installations’ support or digital marketplaces and tourist guides (met in digital and smart city classes).

• SG4: smart health and tele-care services (Healthcare market-driven group) offer distant telematics support to groups of citizens (e.g., elderly people) (appear in digital and smart city classes).

• SG5: smart security services (Public Safety market-driven group) that enhance public safety and emergency (ubiquitous city class).

• SG6: smart environmental services (Utilities market-driven group) address environmental protection and mainly concern waste collection and recycling, emission control, as well as utility services (e.g., energy and water) (met in ubiquitous and eco-city approaches).

• SG7: intelligent transportation (Transportation market-driven group) concern traffic control and public transportation optimization (offered by digital and smart city approaches).

• SG8: typical telecommunication services (Real estate market-driven group) such as broadband connectivity, digital TV etc. (offered by broadband, mobile, digital, smart and ubiquitous approaches).

• SG9: smart education services (Education market-driven group), that concern distant learning services and online libraries (available in smart and digital city approaches). The concept of smart service classification was that when a smart city "migrates" from one class to another, a corresponding change to the offered services is being performed and vice versa. Moreover, the only means to discover in which class a smart city belongs is to investigate literature publications and reports) and the types of services that the city offers (either with person visits or according to the official websites).

The process for the smart city architecture definition consists of the following steps:

• Smart city meta-architecture definition.

• Smart city ICT architecture alternatives' definition.

• Smart city frameworks' and patterns definition.

The above process steps initially result to the definition of the smart city meta-architecture, which incorporates the following components:

• Soft infrastructure: people, knowledge, communities, business processes etc.

• Hard infrastructure: buildings, city facilities (e.g., roads, bridges, telecommunications networks etc.) and utilities (e.g., water, energy, waste, heat etc.)

• ICT-based innovation: both hardware and software solutions, which can be embedded in the above hard and soft infrastructure or deliver corresponding smart services

• Non-ICT based innovation: innovation—beyond the ICT—that addresses smart city dimensions (e.g., creativity, open spaces, recycling and waste management, smart materials, organizational innovation in government, etc.)

• Physical environment: concerns the natural landscape of the city (e.g., ground, forests, rivers, mountains, etc.). In this regard, the resulted multi-tier meta-architecture consists of the following layers from top to bottom

Layer 1 - Natural Environment: respecting all the environmental features where the city is located.

Layer 2 - Hard Infrastructure (Non ICT-based): it contains all the urban facilities (e.g., buildings, roads, bridges, energy-water-waste-heat utilities, etc.).

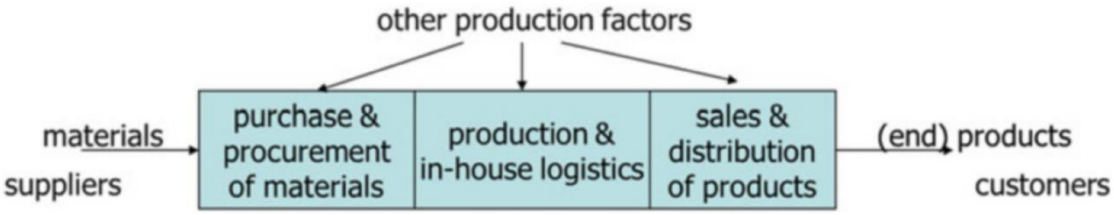
Layer 3 - Hard Infrastructure (ICT-based): it concerns all hardware, with which smart services are being produced and delivered to the end-users (e.g., datacenters, telecommunication networks, IoT, sensors, etc.)

Layer 4 - Smart Services: the smart services that are being offered via both the hard and soft infrastructure (e.g., smart safety, intelligent transportation, smart government, smart water management, etc.).

Layer 5 - Soft Infrastructure: individuals and groups of people living in the city, business process, software applications and data, with which the smart services are executed and being realized.

2 Smart Logistics and transportation

Roughly simplified a company can be considered as a system which receives a specific input, creates goods or services out of it and delivers the output to its customers. These three main activities can be denoted as procurement (or purchasing), production, and distribution (or sales). Very often input and output are physical objects or materials which require a specific handling such as transportation or storage in order to realize the basic functions of a company. These activities are considered the core tasks of logistics



other production factors
materials
suppliers
purchase & procurement of materials
production & in-house logistics
sales & distribution of products
(end) products
customers

Figure 2.1 - The company and its main functions

To be more precise, the input of a company, often called the production factors, can be distinguished into input which is used up during the production and input which is available for a longer period of time. The first group of production factors is often just called materials or consumption factors, whereas the second group includes the classical production factors capital and labor (employees). These factors are also often called resources and include machines, equipment, and buildings. They are not used up during the production but frequently their capacity (available time over a period) needs to be matched with the time needed for production. Also "modern" production factors like information, human capital or management belong to the latter group.

In a more traditional understanding logistics only refers to the materials and could therefore also be called material logistics.

Expressed in a more abstract way, logistics deals with transfers of materials in space, time, and quantity from the procurement of materials needed for production via the storage of materials, intermediate products, and finished products to the physical distribution to customers. Thus, logistics focusses on the planning and execution of spatial, temporal and quantitative transfers.

Spatial transfers could simply be called transportation[24] and can be distinguished into long- and short-distance transports. Long distance transportation means transports between different locations such as warehouses, plants, and different companies that primarily use trucks, trains, ships, and aircrafts. Short-distance transportation means transports inside a location (plant, warehouse). This type of spatial transfer is occasionally also called material flow. In the short-distance transportation usually different devices are used than in long-distance transportation, e.g. fork lifts, conveyors, or automated guided vehicles.

Temporal transfer means “transport” over time, i.e. from today when a material is available to the future when the material is needed. This is the purpose of what we call more simply storage or warehousing.

Quantitative transfers take place when, for instance, large amounts of some goods are provided in smaller quantities. This is one of the usual activities of retailers which buy goods in larger quantities from the producers or wholesalers and usually sell them in smaller amounts to end customers. Changes of quantity also take place when customer orders are fulfilled. Ordered items are picked from the warehouse (where they are usually available in larger quantities), brought together, packaged and sent to the customers.

2.1 Logistics

Council of Logistics Management (1991) defined that logistics is ‘*part of the supply chain process that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information between the point of origin and the point of consumption in order to meet customers’ requirements*’. Johnson and Wood’s definition uses ‘five important key terms’, which are logistics, inbound logistics, materials management, physical distribution, and supply-chain management, to interpret. *Logistics describes the entire process of materials and products moving into, through and out of firm. Inbound logistics covers the movement of material received from suppliers. Materials management describes the movement of materials and components within a firm. Physical distribution refers to the movement of goods outward from the end of the assembly line to the customer. Finally, supply-chain management is somewhat larger than logistics, and it links logistics more directly with the user’s total communications network and with the firm’s engineering staff*[25].

The commonality of the recent definitions is that logistics is a process of moving and handling goods and materials, from the beginning to the end of the production, sale process and waste disposal, to satisfy customers and add business competitiveness. It is ‘*the process of anticipating customer needs and wants; acquiring the capital, materials, people, technologies, and information necessary to meet those needs and wants; optimising the goods- or service-producing network to fulfil customer requests; and utilizing the network to fulfil customer requests in a timely way*’ (Tilanus, 1997). Simply to say, ‘logistics is customer-oriented operation management’.

Logistics services comprise physical activities (e.g. transport, storage) as well as non-physical activities (e.g. supply chain design, selection of contractors, freightage negotiations). Most activities of logistics services are bi-direction. Information systems include modelling and management of decision making, and more important issues are tracking and tracing. It provides essential data and consultation in each step of the interaction among logistics services and the target stations. Infrastructure comprises human resources, financial resources, packaging materials, warehouses, transport and communications. Most fixed capital is for building those infrastructures. They are concrete foundations and basements within logistics systems.

2.1.1 Supply chain management

The trend towards wider definitions of logistics has been incorporated in the idea of what is today called supply chain management. In particular, this concept has been motivated by the awareness that many logistic processes are not just relevant in a considered company but that such processes should also be considered at suppliers and customers for providing a good product or service to end customers. Moreover, from the viewpoint of involved companies, it often makes sense to plan these activities in an integrated way to recover the contribution of the partners in the supply network and to maximize their added value.

The definition of Supply Chain Management is "an integrative philosophy to manage the total flow of a distribution channel from the supplier to the ultimate user". Harland (1996) defines supply chain management as "the management of a network of interconnected businesses involved in the ultimate provision of product and service packages required by end customers". A more official definition comes from the Council of Supply Chain Management Professionals (CSCMP): "Supply chain management encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third party service providers, and customers. In essence, supply chain management integrates supply and demand management within and across companies."

2.1.2 Importance of Logistics

One of the long-term concepts for discussing economic development is the three-sector hypothesis. Roughly speaking, this hypothesis is based on a classification of economic activities into three sectors: The primary sector which deals with the production of raw materials (especially agriculture), the secondary sector which includes manufacturing and industry, and the third or tertiary sector which deals with services. During the maturing of economies usually the following development takes place: First (i.e. some hundred years ago) the primary sector strongly dominates. Then the secondary sector grows and becomes the most important, e.g. during the industrial revolution in Europe and the U.S. in the nineteenth century. Later on, and in particular in the industrialized countries during the second half of the twentieth century, the second sector starts to shrink while the third sector grows and, finally, dominates. These long-term developments demonstrate the importance of logistics because logistic activities can be classified as services and belong, therefore, to the growing third sector.

Although this says little about the growing importance of particular logistics activities, there is further evidence: Along with the long-term economic development, growth and increased standards of living can be observed. These increased standards of living lead to more complex and challenging customer needs. For instance, customers require a higher quality of goods and services, the speed and the security of deliveries have become more important, and special requests like individualized goods or services are often to be taken into account as well. Some of these aspects have clear logistic implications: For a fast and secure supply, good forecasts of demand, an adequate warehousing or fast transportation systems are very important. Other aspects such as the quality of products are partially dependent on a good mastering of logistics. The selection of suppliers, an efficient handling of materials, the avoidance of deficient products, an adequate dealing with customer claims, etc. may contribute to the quality perceived by customers.

Apart from such customer-centric aspects various other global developments in the economy are relevant for the evolution of logistics and supply chain management: One of them is the economic liberalization. Over a longer period of time it has been observed a strengthening of free trade.

A certain consequence of this globalization is that more and more goods are transported over longer distances. As a result, increased transportation costs in the economies can be expected. Over the last few decades, the percentage of the GDP in transportation costs (e.g. for the U.S.) have, as a matter of fact, decreased. This is even more remarkable because the prices for fuel tended to increase over that time (with strong fluctuations). The main reasons for this can be found in aspects relating to technical progress. First of all, many types of vehicles (trucks, planes, vessels) are more fuel-efficient than before. Often, vehicles have become larger (especially ships) which also leads to lower costs per ton transported cargo. Economies of scale do not just arise because of technical reasons. For instance, the crew for a larger ship can be kept constant in size.

Along with such aspects of single transportation activities the whole infrastructure for transports has been upgraded. For instance, turnover activities in a harbor can be done more smoothly today. The container as a well standardized transport equipment plays a major role. A transport chain comprising truck, train or ship transportation (intermodal transport) works more efficiently than in pre-container times. Transportation is nowadays better supported by information and communication technology than in the past. GPS-based navigation systems facilitate the planning and execution of transports. Better planning algorithms allow for more efficient transportation.

Apart from these transportation related aspects, the technical progress in information and communication technologies supports other logistics activities as well (and, of course, business activities in general). Some of the most obvious aspects have been the improved performance of computers and the decrease of respective costs during the last decades. Another obvious development is the rise of the Internet and the emergence of e-business. Let us just consider one example of how logistics is directly influenced by these trends: Because of e-business today goods are more often directly shipped to customers instead of being distributed through retailers. Compared to shipments to retailers this requires a significantly larger number of usually smaller shipments. Another less visible effect of the technological progress is the increased usage of automation technology in companies. Today, companies often master their in-house logistic tasks by using a significant amount of material flow technology or robotics. Examples of such technologies are conveyors for in-house transportation, automatic storage and retrieval systems, automated guided vehicles (AGVs), or an increased usage of mobile devices by humans. One specific technology which has been discussed intensively during the last 10 years is RFID (radio-frequency identification). Small chips (or smart tags) which consist of an electronic circuit (including a memory and possibly sensors) and an antenna for communication can be attached to goods (or parcels). Such technology can be used for a better identification of objects, for a better tracing and tracking during their transportation, or for more intelligent purposes such as the monitoring of a cold chain for frozen food.

Along with new planning technologies and respective software, such technical progress can lead to innovative logistics solutions and/or the decrease of costs.

2.1.3 Smart logistics

The road towards outstanding logistics' performance is Smart Logistics[26]. Smart Logistics equals 3P+I (i.e. Planning, People, Policy and Infrastructure), and is the synchronized interplay of these four key domains. ICT infrastructure is an enabler for planning and scheduling via providing the right information resources at the right time and place.

Smart Logistics is the smart way of the efficient and cost-effective managerial decisions related to the design, planning and control the supply chain Processes. Smart means that planning and scheduling, ICT infrastructure, people and governmental policymaking need to be efficiently aligned.

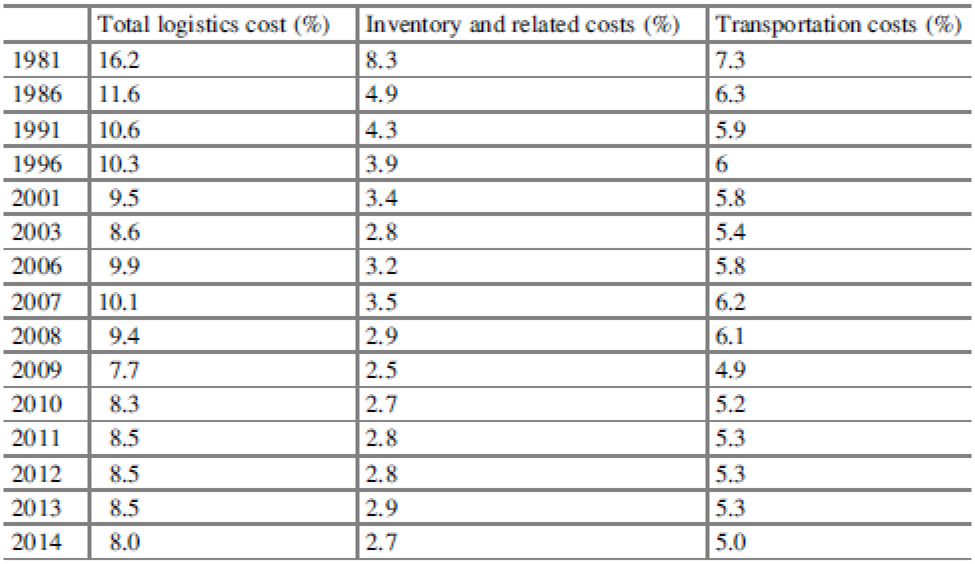
Nowadays, larger quantities along with more detailed and faster information are available. This allows for better planning and scheduling. But this is also a challenge as many planning and scheduling tools are not able to handle this amount and quality of information.

2.2 Transportation

The field of transportation belongs to the most important areas of logistics. Transportation problems occur in various variations and complexity and require a careful planning due to their significance and the need to solve them efficiently, i.e. with high quality and low costs. In general, transportation planning requires to determine the kind and quantity of the goods that are shipped, from where to where, at which time and by using which resources, in particular by which vehicles. The determination of a specific path or route belongs to the field of transportation planning as well.

Transportation takes a crucial part in the manipulation of logistic. For industries, logistics helps to optimise the existing production and distribution processes based on the same resources through management techniques for promoting the efficiency and competitiveness of enterprises. The key element in a logistics chain is transportation system, which joints the separated activities. Logistics and Transportation costs and transportation systems influence the performance of logistics system hugely.

Table 2.1 shows the logistics costs for the US expressed in percentage of the gross domestic product (GDP) of this country. Total logistics costs and the two main subcategories, inventory-related costs and transportation-related costs, are depicted (The category administrative costs are omitted so that the two percentages of the subcategories are slightly less than the percentage for the total logistics costs.) Moreover, the developments of these costs from the early 1980s until 2012 are shown.



	Total logistics cost (%)	Inventory and related costs (%)	Transportation costs (%)
1981	16.2	8.3	7.3
1986	11.6	4.9	6.3
1991	10.6	4.3	5.9
1996	10.3	3.9	6
2001	9.5	3.4	5.8
2003	8.6	2.8	5.4
2006	9.9	3.2	5.8
2007	10.1	3.5	6.2
2008	9.4	2.9	6.1
2009	7.7	2.5	4.9
2010	8.3	2.7	5.2
2011	8.5	2.8	5.3
2012	8.5	2.8	5.3
2013	8.5	2.9	5.3
2014	8.0	2.7	5.0

Table 2.1 - Logistics costs in the US

Transporting is required in the whole production procedures, from manufacturing to delivery to the final consumers and returns. Only a good coordination between each component would bring the benefits to a maximum.

A strong system needs a clear frame of logistics and a proper transport implements and techniques to link the producing procedures.

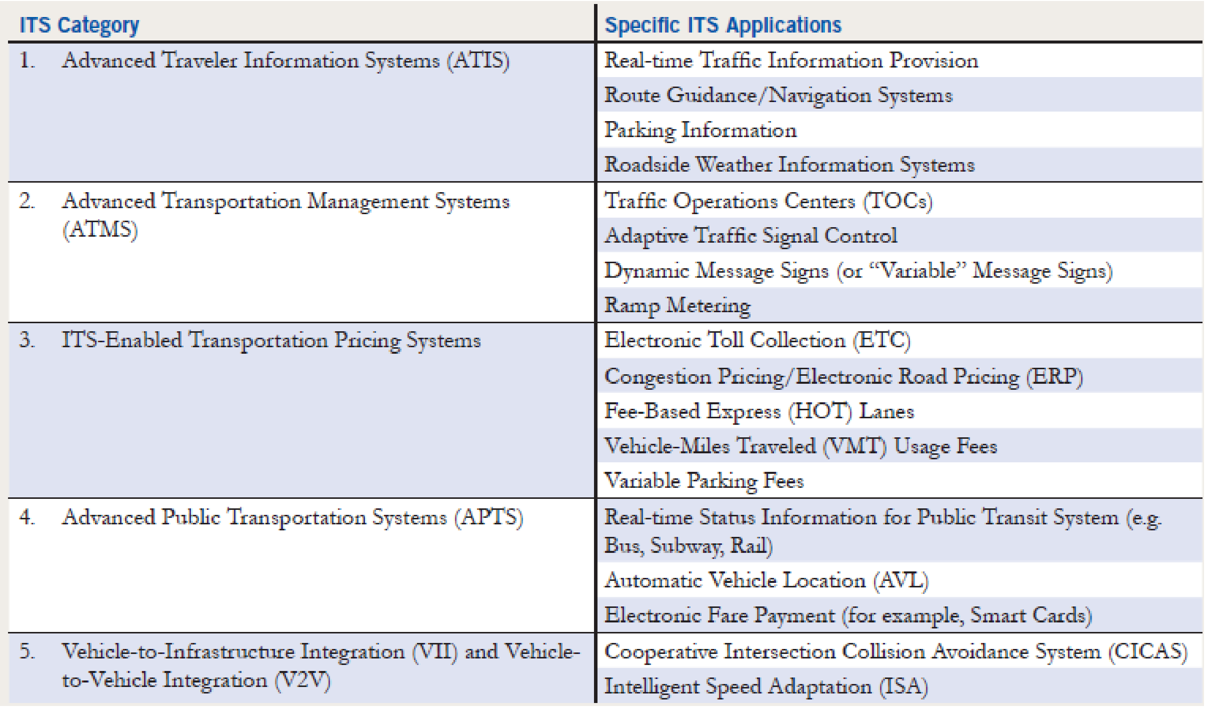
2.2.1 Intelligent transportation systems

The operation of transportation determines the efficiency of moving products. The progress in techniques and management principles improves the moving load, delivery speed, service quality, operation costs, the usage of facilities and energy saving.

Smart transportation also labeled as Intelligent Transportation Systems (ITS) combine information, telecommunications, positioning and automation technologies in an attempt to improve transportation's safety, efficiency within the urban space and reduce corresponding environmental impact (ITU 2014b, 2007).

Intelligent transportation systems[27] include a wide and growing suite of technologies and applications. ITS applications can be grouped within five summary categories: 1) Advanced Traveler Information Systems provide drivers with real-time information, such as transit routes and schedules; navigation directions; and information about delays due to congestion, accidents, weather conditions, or road repair work. 2) Advanced Transportation Management Systems include traffic control devices, such as traffic signals, ramp meters, variable message signs, and traffic operations centers. 3) ITS-Enabled Transportation Pricing Systems include systems such as electronic toll collection (ETC), congestion pricing, fee-based express (HOT) lanes, and vehicle miles traveled (VMT) usage-based fee systems. 4) Advanced Public Transportation Systems, for example, allow trains and buses to report their position so passengers can be informed of their real-time status (arrival and departure information). 5) Fully integrated intelligent transportation systems, such as vehicle-to-infrastructure (VII) and vehicle-to-vehicle (V2V) integration, enable communication among assets in the transportation system, for example, from vehicles to roadside sensors, traffic lights, and other vehicles.

Given the wide range of intelligent transportation systems, it is useful to organize discussion of ITS applications through a taxonomy that arranges them by their primary functional intent (with the acknowledgment that many ITS applications can serve multiple functions or purposes). While this list is not inclusive of all possible ITS applications, it includes the most prominent ones, which are the focus of this report (Table 2.2). ITS applications can be grouped within five primary categories: Advanced Traveler Information Systems (ATIS), Advanced Transportation Management Systems (ATMS), ITS-Enabled Transportation Pricing Systems, Advanced Public Transportation Systems (APTS), and Fully Integrated ITS Systems (VII and V2V Systems).



ITS Category	Specific ITS Applications
1. Advanced Traveler Information Systems (ATIS)	Real-time Traffic Information Provision
	Route Guidance/Navigation Systems
	Parking Information
	Roadside Weather Information Systems
2. Advanced Transportation Management Systems (ATMS)	Traffic Operations Centers (TOCs)
	Adaptive Traffic Signal Control
	Dynamic Message Signs (or "Variable" Message Signs)
	Ramp Metering
3. ITS-Enabled Transportation Pricing Systems	Electronic Toll Collection (ETC)
	Congestion Pricing/Electronic Road Pricing (ERP)
	Fee-Based Express (HOT) Lanes
	Vehicle-Miles Traveled (VMT) Usage Fees
	Variable Parking Fees
4. Advanced Public Transportation Systems (APTS)	Real-time Status Information for Public Transit System (e.g. Bus, Subway, Rail)
	Automatic Vehicle Location (AVL)
	Electronic Fare Payment (for example, Smart Cards)
5. Vehicle-to-Infrastructure Integration (VII) and Vehicle-to-Vehicle Integration (V2V)	Cooperative Intersection Collision Avoidance System (CICAS)
	Intelligent Speed Adaptation (ISA)

Table 2.2 - ***Classifying Contactless Mobile Payments Applications***

2.2.2 ITS Applications: Definitions and Technologies

i) Advanced Traveler Information Systems

Perhaps the most-recognized ITS applications, Advanced Traveler Information Systems (ATIS) provide drivers with real-time travel and traffic information, such as transit routes and schedules; navigation directions; and information about delays due to congestion, accidents, weather conditions, or road repair work. The most effective traveler information systems are able to inform drivers in real-time of their precise location, inform them of current traffic or road conditions on their and surrounding roadways, and empower them with optimal route selection and navigation instructions, ideally making this information available on multiple platforms, both in-vehicle and out.

ii) Advanced Transportation Management Systems

Advanced Transportation Management Systems (ATMS) include ITS applications that focus on traffic control devices, such as traffic signals, ramp metering, and the dynamic (or "variable") message signs on highways that provide drivers real-time messaging about traffic or highway status. Traffic Operations Centers (TOCs), centralized traffic management centers run by cities and states worldwide, rely on information technologies to connect sensors and roadside equipment, vehicle probes, cameras, message signs, and other devices together to create an integrated view of traffic flow and to detect accidents, dangerous weather events, or other roadway hazards.

iii) ITS-Enabled Transportation Pricing Systems

ITS have a central role to play in funding countries' transportation systems. The most common application is electronic toll collection (ETC), also commonly known internationally as "road user charging," through which drivers can pay tolls automatically via a DSRC-enabled on-board device or tag placed on the windshield.

iv) Advanced Public Transportation Systems

Advanced Public Transportation Systems (APTS) include applications such as automatic vehicle location (AVL), which enable transit vehicles, whether bus or rail, to report their current location, making it possible for traffic operations managers to construct a real-time view of the status of all assets in the public transportation system. APTS help to make public transport a more attractive option for commuters by giving them enhanced visibility into the arrival and departure status (and overall timeliness) of buses and trains. This category also includes electronic fare payment systems for public transportation systems, which enable transit users to pay fares contactlessly from their smart cards or mobile phones using near field communications technology. Advanced public transportation systems, particularly providing “next bus” or “next train information, are increasingly common worldwide.

v) Vehicle-to-infrastructure Integration (VII) and Vehicle-to-vehicle (V2V) Integration

Vehicle-to-infrastructure integration is the archetype for a comprehensively integrated intelligent transportation system. In the United States, the objective of the VII Initiative—as of January 2009 rebranded as IntelliDriveSM—has been to deploy and enable a communications infrastructure that supports vehicle-to-infrastructure, as well as vehicle-to-vehicle, communications for a variety of vehicle safety applications and transportation operations. IntelliDrive envisions that DSRC-enabled tags or sensors, if widely deployed in vehicles, highways, and in roadside or intersection equipment, would enable the core elements of the transportation system to intelligently communicate with one another, delivering a wide range of benefits. For example, IntelliDrive could enable cooperative intersection collision avoidance systems (CICAS) in which two (or more) DSRC-equipped vehicles at an intersection would be in continuous communication either with each other or with roadside devices that could recognize when a collision between the vehicles appeared imminent (based on the vehicles’ speeds and trajectories) and would warn the drivers of an impending collision or even communicate directly with the vehicles to brake them. IntelliDrive, by combining both vehicle-to-vehicle and vehicle-to-infrastructure integration into a consolidated platform, would enable a number of additional ITS applications, including adaptive signal timing, dynamic re-routing of traffic through variable message signs, lane departure warnings, curve speed warnings, and automatic detection of roadway hazards, such as potholes, or weather-related conditions, such as icing.

Another application enabled by vehicle-to-infrastructure integration is intelligent speed adaptation (ISA), which aims to assist drivers in keeping within the speed limit by correlating information about the vehicle’s position (for example, through GPS) with a digital speed limit map, thus enabling the vehicle to recognize if it is exceeding the posted speed limit. The system could either warn the driver to slow down or be designed to automatically slow the vehicle through automatic intervention. France is currently testing deployment of an ISA system that would automatically slow fast-moving vehicles in extreme weather conditions, such as blizzards or icing.

2.2.3 Benefits of Intelligent Transportation Systems

Applying information technology to a country’s transportation network delivers five key classes of benefits by: 1) increasing driver and pedestrian safety, 2) improving the operational performance of the transportation network, particularly by reducing congestion, 3) enhancing personal mobility and convenience, 4) delivering environmental benefits, and 5) boosting productivity and expanding economic and employment growth.

3 E-Commerce as a critical part of smart logistics operations

In Chapter three we make an introduction to e-commerce because we want to explain the importance of logistics and big data on e-commerce. We conclude into some specific requirements which revealed numerous areas that potential big data contribution and implementation could bring significant improvements in terms of cost and customer service.

3.1 An Introduction to Electronic Commerce

Electronic commerce is a powerful concept and process that has fundamentally changed the current of human life. Electronic commerce is one of the main criteria of revolution of Information Technology and communication in the field of economy. This style of trading due to the enormous benefits for human has spread rapidly. Certainly, can be claimed that electronic commerce is canceled many of the limitations of traditional business. For example, form and appearance of traditional business has fundamentally changed. These changes are basis for any decision in the economy. Existence of virtual markets, passages and stores that have not occupy any physical space, allowing access and circulation in these markets for a moment and anywhere in the world without leaving home is possible. Select and order goods that are placed in virtual shop windows at unspecified parts of the world and also are advertising on virtual networks and payment is provided through electronic services, all of these options have been caused that electronic commerce is considered the miracle of our century.

E-commerce growth

Digital technology changes the way consumers shop and the way consumers wish to receive their purchases. Shopping habits have changed rapidly during the last decade and a high percentage of consumers now shop online, following the spread of IT systems such as laptops, tablets and smartphones. Today, 57% of European Internet[28] users shop online. The e-commerce sector is booming as almost all growth in retail comes from e-commerce. The European B2C e-commerce sales have been growing steadily since 2011. Still, the growth rate has decreased the last few years, from 18.4% in 2011 to 13.3% in 2015. This trend is expected to continue in 2016, as a growth rate of 12.0% has been forecasted, resulting in a European B2C ecommerce turnover of €509.9bn as shown in Figure 3.1 and 3.2 Greece is included among the TOP 10 countries in terms of B2C e-Commerce growth rate with a growth rate of 18.8%.

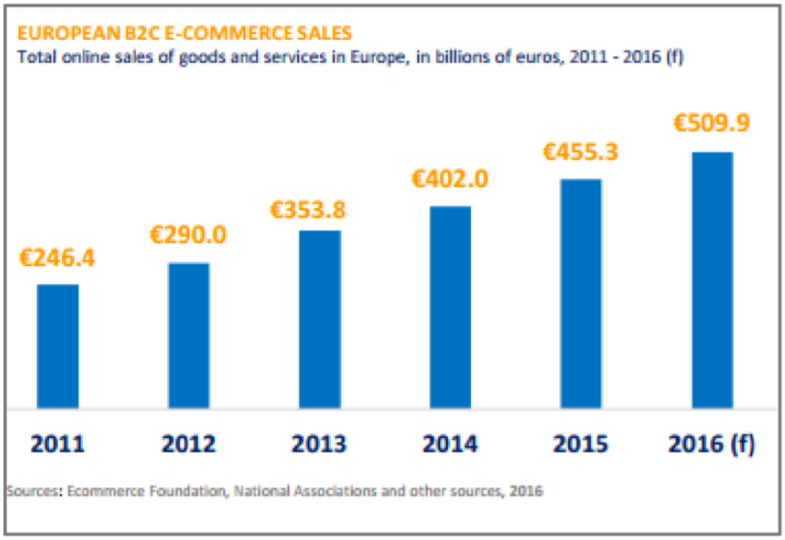


Figure 3.1 - ***European B2C e-Commerce Sales***

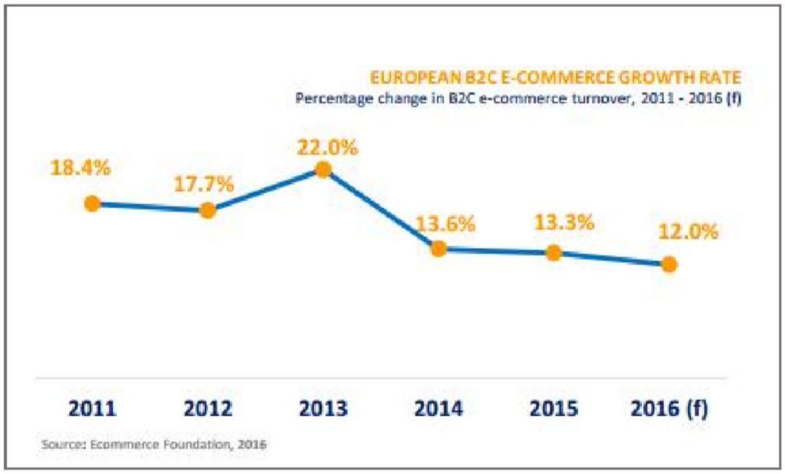


Figure 3.2 - ***European B2C e-Commerce Growth Rate***

However, the full potential of the European e-commerce market has not yet been reached as the e-commerce penetration can still be increased; only 16% of SMEs sell online – and less than half of those sell online across borders (7.5%). According to a recent expert study[29], the driver that is currently most influential and is having most impact on the retail scene is shoppers' convenience expectation: a very demanding issue in an increasingly complex retail and shopping environment[29].

Moreover, customers today are demanding a high level of service from retailers, regardless of how they buy products or receive delivery. The omni-channel growth has created new challenges for the retailers. Whether their business is online, in physical stores, or both, all retailers face similar challenges. They have to deliver a seamless customer experience at every touch point, maximize sales across every channel and device, and live up to their promises regarding product availability and delivery. Taking a look at the UK and Germany, we see that more than 65% of shoppers aged 21 or younger prefer to have both a physical and an online experience. Consumers expect shopping to be convenient, and expectations are still rising: 40% are looking for even easier shopping across on- and offline channels, 41% expect improved customer service, 45% want improved delivery service, and 46% ask for easier return and refund[28].

The outcomes of the annual Greek e-Commerce survey, which ELTRUN[30], conducts annually reinforce the above ones. The survey was conducted in the beginning of 2016 and almost 1000 questionnaires were collected. The survey insights revealed the problems that customers deal with at the delivery process and highlighted the main customers' requirements concerning the delivery process. Indicatively, only 80% declared that they received the product in the right condition. Moreover, only 26% of the consumers recognize that the total procedure of delivery is simple and only 23% declares that e-shops can react properly in urgent situations. Order traceability is recognized as one of the major requirements of the delivery process, since a significant part of the consumers cannot track their order. Therefore, online retailers and logistics companies should meet the e-commerce customers' requirements of more qualitative, traceable delivery processes.

Responding to the main emerging trend triggered by e-commerce growth, the transportation industry is undergoing a major transformation nowadays. The e-commerce trend has created a dynamic and turbulent environment, where the density of the distribution network consisted by many delivery points, multiple delivery channels and last-mile delivery requirements significantly increases the distribution cost. Therefore, the reduction of delivery costs can become a significant competitive advantage. In this context, complexity has also raised by the variety of distribution requests, e.g. orders and returns from various consumers at different points. Therefore, e-commerce logistics turned to be a key area of innovation and one of the European Commission's key policy areas. Retail formats that deliver on this complexity will therefore be more successful in the future[28].

E-commerce for physical goods generates a significant demand for dedicated delivery services that results in increasing fragmentation of shipments in the "last mile"[31,32]. In particular, home delivery services, which are usually preferred by the online consumers, contribute to the atomization of parcel flows thus causing particular problems within the urban areas. However, alternative delivery solutions such as pick-up points and lockers are growing fast, especially in urban areas. Stores are becoming fulfillment centers, serving as pick-up locations for online orders and customers require more dynamic delivery services for the last mile rendering this process a challenging task.

As consumer turns to be an important stakeholder in the e-commerce area, a consumers' survey in the beginning of 2016 focusing on e-Commerce deliveries in Greece was conducted to get more insights about the alternative delivery methods consumers prefer. Home delivery is by far the most preferable delivery method for consumers according to the survey results (80%), but, at the same time, participants show a positive attitude in using potential alternative delivery methods, such as pick up points and that the solution of pick up points would be more preferable if it was offered by more e-shops[30].

3.2 Big data & e-commerce logistics

Many studies show that customers consider the logistics performance as an important factor of E-commerce[33,34]. Distribution of goods to customers and in particular, last mile operation is probably the most demanding process in e-commerce. This is because in e-commerce customer distribution deals with frequent personalized orders resulting in high costs and difficult to manage processes. Outbound delivery costs are high because customer orders do not fill a truck resulting in low utilization of delivery trucks and personnel and longer routes. In addition, re-distribution, or returns system, is considered to be one of the most problematic and costly activities in e-commerce logistics. An efficiently organized return process, as part of the distribution system, can help to retain the customer by minimizing his inconvenience and reducing pick up and return handling costs.

Logistics is considered to be a basic part of e-commerce[35] and, according to literature findings, it is undoubtedly an area where big data can be extremely fertile (Swaminathan, 2012). Big data collection and analysis can make activities in e-commerce logistics much more efficient[36]. Personalized services, dynamic pricing, predictive analytics, supply chain optimization and visibility are some of the core big data application areas in these fields[36,37,38]. In the e-commerce context, big data can be used to improve decision making in all activities involving infrastructure and operation on one hand, and consumers' behavior on the other, thus archiving a better matching between supply and demand.

Exploitation of big data in logistics requires matching data sources, analytic procedures and business understanding. Three main areas of logistic applications, where big data analytics can provide effective decision support, are identified in the literature (Figure 3.3): operational efficiency and network planning, customer experience and new business models[39.]

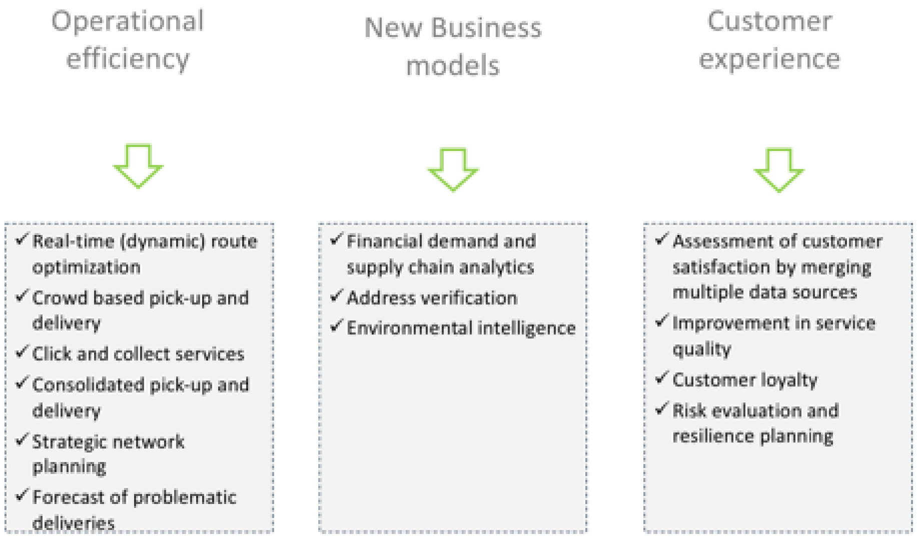

Operational efficiency
New Business models
Customer experience
✓Real-time (dynamic) route optimization
✓Crowd based pick-up and delivery
✓Click and collect services
✓Consolidated pick-up and delivery
✓Strategic network planning
✓Forecast of problematic deliveries
✓Financial demand and supply chain analytics
✓Address verification
✓Environmental intelligence
✓Assessment of customer satisfaction by merging multiple data sources
✓Improvement in service quality
✓Customer loyalty
✓Risk evaluation and resilience planning

Figure 3.3 - Big data applications in logistics

Operational efficiency: Using big data analytics, there is a huge potential for improving operational efficiency. Operational efficiency includes both last mile optimization, as well as network planning. Last mile efficiency can be considerably improved by real time optimization and consolidated pick-up and delivery. A second area of big data applications in logistics regarding operational efficiency field is network planning, which includes decisions concerning warehouses, distribution centers and custom-built vehicles.

Business Models: Using the huge amount of data, logistics providers can develop new business models and provide additional services[39]. New business models include financial demand and supply chain analytics, address verification and correct geocoding. All these applications call for big data analysis methodological approach and can help enterprises gain additional revenue sources.

Customer experience: Big data analytics help to maximize customer satisfaction and understand customer demand[40]. Use cases include both customer value management and supply chain risk management[39]. Merging data from multiple sources (sales information, data derived from social media, discussion forums and e-commerce catalogues) valuable information for effective customer relation management can be obtained that results in assessing customer satisfaction, improving service quality and increasing customer loyalty. Moreover, the use of big data analytics can help towards capturing consumer behavior and restricting supply chain risk management.

Operational Efficiency - Last mile optimization cases

Real-time (dynamic) route optimization: Last mile is the most expensive distribution leg[41] and it can be considerably improved if we use real time information from different sources (traffic data / sensors / real time events) to dynamically optimize routing and provide drivers with directions on the spot[42,43]. In e-commerce deliveries, dynamic routing is even more challenging due to the fact that orders change unpredictably and dynamically[44]. Dynamic Vehicle Routing allows for taking into account both the delivery and the new return requests. As indicated earlier, re-distribution or Returns system is considered to be one of the most problematic and costly activities in e-commerce logistics and thus, dynamically rerouting by combining deliveries and new return requests in real time could improve operational efficiency.

***Forecast of failed or problematic deliveries**:* Problematic deliveries are also recognized as one of the major problems in e-commerce delivery since they contribute to increased costs, emissions, and wasted time and effort[45]. Delivery efficiency[46] can be improved by identifying problematic zip code areas and applying changes in last mile delivery regarding time, location and route.

***Crowd based pick-up and delivery**:* An additional application that optimizes last mile delivery is crowd –based pick up and delivery that employs different classes of commuters and taxi drivers to undertake paid deliveries[39]. These operations are based on real time data streams involving event processing and geo-correlation big data techniques[47].

***Consolidated pick-up and delivery**:* Based on recent professional reports and academic publications[48,49], delivery efficiency can be considerably improved by combining routes of multiple supply chain stakeholders (e.g. e-commerce retail players, 3PLs). Consolidating routes from different companies can help in making more efficient the routing process and also increase the utilization capacity of vehicles. This requires exploiting and consolidating information from multiple data sources. This use case appears to be a novel feature in the practice of collaborative logistics in the context of e-commerce deliveries.

Operational Efficiency - Network and Operational planning Cases

Network planning: Big data techniques support network planning by analyzing historic data, seasonal trends, emerging freight flows as well as more general information concerning regional growth. Accurate long- term demand forecasts based on detailed data are generated in order to improve strategic decisions and support investments concerning network planning such as location of distribution points and warehouses. Demand fluctuations or uncertainties constitute are the most challenging problem that network design has to face[50].

Dynamic inventory routing: Inventory deliveries which are based on sales data in a Vendor Managed Inventory (VMI) framework can benefit from real time dynamic routing. In this case, routing is organized on a short-term basis. Special purpose models can provide considerable help in improving the routing schedule, increasing the precision of the future demand estimations and decreasing the requirement of extra (non-prescheduled) vehicle routes and their respective costs[50,51].

Location of click and collect points or pick up points: Click and collect service is an alternative delivery method that can offer flexibility and cost reduction and thus increase customers' satisfaction. Arguing that pick up points appear to be a novel way for limiting the costs of last mile e-commerce delivery[52]. According to institute report[53] the future of online grocery in Europe is closely related to successful operation of the click and collect service. This model gives retailers easier entry into the online-grocery space, since it has much less daunting economics than home-delivery service. In the case of e-commerce, adopting a hybrid system of retail stores and click and collect points by adding click and collect points to an existing distribution network on one hand offers an additional service to the consumers, but on the other entails the risk of creating points of unutilized capacity. For these reasons making the right mix of physical retail stores and click and collect points is a strategic decision concerning the distribution network that should be carefully evaluated[54]. Location of click and collect points needs to take into account both transportation flows and strategic planning of the distribution centers. There is no doubt that the development of convenient pickup points is crucial for the successful implementation of such a service and has attracted attention in e-commerce delivery processes[52]. According to customer density appears to be a crucial factor regarding the correct location of these points[55].

In summary, Big Data Analytics (BDA) play an essential role in improving the efficiency and effectiveness of E-commerce logistics[56]. Although there are some existing industrial business use cases, implementation of big data techniques appears to be an untapped asset across industries[57]. Based on the existing literature, one of the most problematic areas of supply chain logistics in general and e-commerce logistics in particular, is distribution of goods because it is mainly performed on a personalized basis, resulting in high costs and difficult to manage processes. Outbound delivery costs are high because customer orders do not fill a truck resulting in low utilization of delivery trucks and personnel and longer routes. Future logistics models should focus on providing high service quality in last mile distribution and improving operational efficiency for companies[56].

Literature review findings indicate that some effective ways for improving efficiency in e-commerce distribution are: providing alternative methods of delivery and augmenting the distribution network with click and collect points, forecasting problematic deliveries, implementing efficient returns management and routing, combining deliveries and returns in real time. Moreover, there is room for action in the field of shared logistics in e-commerce (consolidating distribution by arranging cooperation of different logistics providers). Although **shared logistics** have been used in some cases, this type of collaboration has not yet been implemented in e-commerce logistics and needs further investigation.

Last mile delivery is recognized as the most expensive part of the delivery process. The last mile delivery cost is also increased by the phenomenon of "cash on delivery" sometimes called "collect on delivery**"**. Cash on delivery is the sale of goods by mail order where payment is made on delivery rather than in advance. If the goods are not paid for, they are returned to the retailer. Cash on delivery is a payment method which gives the consumer the right not to accept the delivery. Therefore, except for the cost of "cash on delivery" option, the delivery cost is also increased by the number of undelivered orders that needs to be returned. The last mile delivery is also recognized as the most challenging one as it directly affects the consumers' perceptions about a brand. For example, companies declare that most of customers' complaints are related to the wide time window delivery that forces them to wait at home until they receive the product.

Current delivery options include home deliveries, deliveries at physical stores or couriers offices and deliveries at pick up points e.g. gas stations. Retailers investigate the possibility of alternative delivery options and stressed the fact that the new big data technologies that can collect data from various partners, from various sources (e.g. social media data, sales data, locations data and geolocation data) can provide solutions that can decrease the delivery cost and also improve the efficiency of the distribution process.

Another important aspect in the procedure of e-commerce logistics is the "reverse logistics" as the European law gives to customers the opportunity to return the product within 14 days from the order date. The total return percentage varies per industry.

A major problem that many 3PLs companies also face is related to the reverse logistics and the empty runs. Even if the load factor of a vehicle is high and very satisfying, the same vehicle returns without any goods at the end of the day, and this is translated only to costs (regardless of whether it is paid by the companies themselves or not). Therefore, collaboration in the returns area was recognized as a potential scenario of applying big data solutions.

Couriers play a significant role in the e-commerce logistics chain as they are responsible for the last mile distribution and the respective cost. Moreover, they can provide traceability data of their orders that can support a series of advanced analytics in e-commerce logistics and can release a series of dynamic delivery services that can be used by consumers.

The following services can be supported by a service that exploits logistics data and provide a service where a customer can record its new requests (change of delivery point, return requests) and then modify the routing schedule:

- Use of POS by couriers in order to replace the cash on delivery method.
- Sending SMS, which gives to customers the opportunity to change the place of delivery or the time of delivery.
- Sending SMS, which gives to customers the opportunity to change at the same time the place and the time of delivery.
- A unified system which endorses the procedure of reverse logistics.

Moreover, as alternative solutions for reducing the cost came out the following:

(1) cooperation between different warehouses, especially in non-urban areas,

(2) cooperation between couriers (a solution which takes into consideration the legal framework of every country),

(3) cooperation between e-shops in order to share the same pick-up points/lockers. However, the market has still an immature collaboration culture and the high competition make some companies resistant to collaborate even if they recognize a potential value.

3.3 Key outcomes and pilot requirements

After an extensive literature review on big data, e-commerce, logistics, supply chain areas and delivery policies, we present the general requirements as they came up. We elaborated on them and we conclude here with some more specific requirements. Requirement analysis revealed numerous areas that potential big data contribution and implementation could bring significant improvements in terms of cost and customer service. In the following sections, we summarize the main requirements that were revealed during requirements analysis stage.

▪ **Req. 1: To explore how collaboration could be applied in the e-commerce logistics domain in order to address the current challenges and support firms to decrease cost and improve the overall distribution process performance**. Collaboration among supply chain stakeholders appeared to be one of the most challenging facts in the context of e-commerce logistics. Last mile delivery challenges, the demand for more responsive supply chain and the need of cost reduction have brought collaborative approaches in the forefront. For example, consolidated pick-up and delivery process by combining routes of multiple supply chain stakeholders[48,49] can contribute in operational efficiency in the last mile cases. Even though the potential collaboration could be the key to delivery cost reduction, companies in e-commerce are reluctant to share data and common vehicles as came out during the interviews and they need to have more evidence about the value of the collaboration.

Req. 2: To identify potential synergies among the e-commerce stakeholders at the reverse logistics. Reverse logistics is considered to be one of the most complicated and costly process in e-commerce logistics according both to the literature and the industry interview outcomes. Costs and the average percentage of returns vary significantly between different industries (clothes vs electronics) and it reaches up to 30% in the case of fashion industry.

▪ **Req. 3: To analyse current distribution processes in order to depict the current distribution patterns and forecast future problems.** The abundance of data (routing data, orders, customer data) in association with big data analysis approach can dynamically contribute to identifying patterns of distribution, enhance the visualization of the distribution processes and also support the identification of the problematic processes and the forecasting of failed de-liveries and returns, thus improving operational efficiency.

▪ **Req. 4: To provide alternative shipping methods at the consumer in order to increase customers' satisfaction and provide lower prices.** Alternative last mile delivery methods such as click & collect and pick up points could provide both lower costs to the retailers, 3PL and courier companies additionally to better customer experience and service to the customers. Locating click & collect and pick up points is a critical requirement in the context of last mile optimization.

▪ **Req. 5: To enable dynamically changing supply chains that take into consideration various routing and customer preference characteristics in order to decrease logistics costs and to increase customers' satisfaction.** It is identified that customers' need to dynamically change the time and the location of the delivery and Courier Companies are already working on future services to this direction. One of the most challenging requirements on the e-Commerce logistics is the real time dynamic route optimization where real time information from different sources (traffic data/sensors/real time events) is used in or-der to dynamically optimize routing[42,43].

Req. 6: To identify users' problems relevant to the logistics procedure, via the analysis of social media data. Social media growth reveals the opportunity to exploit the open data provided through them. Online consumers express their opinion (negative or positive) via comments, either in blogs/forums or in social networking sites. These data can be analysed to provide insights and identify problems that are relevant to the logistics procedure. Algorithms based on text mining and meaning techniques could be used in order to interpret the data and extract the proper information.

4 Technologies for optimization of smart logistics operations: Big data mining, analytics & data classification

In chapter four we present general techniques and algorithms of text mining, text classification and text clustering as in Chapter 5 based on text mining and meaning techniques we identify users' problems relevant to the logistics procedure, via the analysis of social media data.

4.1 An introduction to Text Mining

Data mining is a field which has seen rapid advances in recent years because of the immense advances in hardware and software technology which has led to the availability of different kinds of data. This is particularly true for the case of text data, where the development of hardware and software platforms for the web and social networks has enabled the rapid creation of large repositories of different kinds of data. In particular, the web is a technological enabler which encourages the creation of a large amount of text content by different users in a form which is easy to store and process. The increasing amounts of text data available from different applications has created a need for advances in algorithmic design which can learn interesting patterns from the data in a dynamic and scalable way.

While structured data is generally managed with a database system, text data is typically managed via a search engine due to the lack of structures. A search engine enables a user to find useful information from a collection conveniently with a keyword query, and how to improve the effectiveness and efficiency of a search engine has been a central research topic in the field of information retrieval, where many related topics to search such as text clustering, text categorization, summarization, and recommender systems are also studied.

However, research in information retrieval has traditionally focused more on facilitating information access rather than analyzing information to discover patterns, which is the primary goal of text mining. The goal of information access is to connect the right information with the right users at the right time with less emphasis on processing or transformation of text information. Text mining can be regarded as going beyond information access to further help users analyze and digest information and facilitate decision making. There are also many applications of text mining where the primary goal is to analyze and discover any interesting pattterns, including trends and outliers, in text data, and the notion of a query is not essential or even relevant.

A number of key characteristics distinguish text data from other forms of data such as relational or quantitative data. This naturally affects the mining techniques which can be used for such data. The most important characteristic of text data is that it is sparse and high dimensional. For example, a given corpus may be drawn from a lexicon of about 100,000 words, but a given text document may contain only a few hundred words.

Thus, a corpus of text documents can be represented as a sparse term document matrix of size n×d, when n is the number of documents, and d is the size of the lexicon vocabulary. The (i, j)th entry of this matrix is the (normalized) frequency of the jth word in the lexicon in document i. The large size and the sparsity of the matrix has immediate implications for a number of data analytical techniques such as dimensionality reduction. In such cases, the methods for reduction should be specifically designed while taking this characteristic of text data into account. The variation in word frequencies and document lengths also lead to a number of issues involving document representation and normalization, which are critical for text mining.

Furthermore, text data can be analyzed at different levels of representation. For example, text data can easily be treated as a bag-of-words, or it can be treated as a string of words. However, in most applications, it would be desirable to represent text information semantically so that more meaningful analysis and mining can be done. For example, representing text data at the level of named entities such as people, organizations, and locations, and their relations may enable discovery of more interesting patterns than representing text as a bag of words. Unfortunately, the state of the art methods in natural language processing are still not robust enough to work well in unrestricted text domains to generate accurate semantic representation of text. Thus, most text mining approaches currently still rely on the shallower word-based representations, especially the bag-of-words approach, which, while losing the positioning information in the words, is generally much simpler to deal with from an algorithmic point of view than the string-based approach. In special domains (e.g., biomedical domain) and for special mining tasks (e.g., extraction of knowledge from the Web), natural language processing techniques, especially information extraction, are also playing an important role in obtaining a semantically more meaningful representation of text. Recently, there has been rapid growth of text data in the context of different web-based applications such as social media, which often occur in the context of multimedia or other heterogeneous data domains. Therefore, a number of techniques have recently been designed for the joint mining of text data in the context of these different kinds of data domains. For example, the Web contains text and image data which are often intimately connected to each other and these links can be used to improve the learning process from one domain to another. Similarly, cross-lingual linkages between documents of different languages can also be used in order to transfer knowledge from one language domain to another. This is closely related to the problem of transfer learning.

Algorithms for Text Mining

In this section, we will explore the key problems arising in the context of text mining. We will also present the organization of the different chapters of this book in the context of these different problems. We intentionally leave the definition of the concept ”text mining” vague to broadly cover a large set of related topics and algorithms for text analysis, spanning many different communities, including natural language processing, information retrieval, data mining, machine learning, and many application domains such as the World Wide Web and Biomedical Science. We have also intentionally allowed (sometimes significant) overlaps between chapters to allow each chapter to be relatively self contained, thus useful as a standing-alone chapter for learning about a specific topic.

Information Extraction from Text Data: Information Extraction is one of the key problems of text mining, which serves as a starting point for many text mining algorithms. For example, extraction of entities and their relations from text can reveal more meaningful semantic information in text data than a simple bag-of-words representation and is generally needed to support inferences about knowledge buried in text data.

Text Summarization: Another common function needed in many text mining applications is to summarize the text documents in order to obtain a brief overview of a large text document or a set of documents on a topic. Summarization techniques generally fall into two categories. In extractive summarization, a summary consists of information units extracted from the original text; in contrast, in abstractive summarization, a summary may contain “synthesized” information units that may not necessarily occur in the text documents. Most existing summarization methods are extractive, and in Chapter 3, we give a brief survey of these commonly used summarization methods.

Unsupervised Learning Methods from Text Data: Unsupervised learning methods do not require any training data, thus can be applied to any text data without requiring any manual effort. The two main unsupervised learning methods commonly used in the context of text data are clustering and topic modeling. The problem of clustering is that of segmenting a corpus of documents into partitions, each corresponding to a topical cluster. The problems of clustering and topic modeling are closely related. In topic modeling we use a probabilistic model in order to determine a soft clustering, in which each document has a membership probability of the cluster, as opposed to a hard segmentation of the documents. Topic models can be considered as the process of clustering with a generative probabilistic model. Each topic can be considered a probability distribution over words, with the representative words having the highest probability. Each document can be expressed as a probabilistic combination of these different topics. Thus, a topic can be considered to be analogous to a cluster, and the membership of a document to a cluster is probabilistic in nature. This also leads to a more elegant cluster membership representation in cases in which the document is known to contain distinct topics. In the case of hard clustering, it is sometimes challenging to assign a document to a single cluster in such cases. Furthermore, topic modeling relates elegantly to the dimension reduction problem, where each topic provides a conceptual dimension, and the documents may be represented as a linear probabilistic combination of these different topics. Thus, topic-modeling provides an extremely general framework, which relates to both the clustering and dimension reduction problems.

LSI and Dimensionality Reduction for Text Mining: The problem of dimensionality reduction is widely studied in the database literature as a method for representing the underlying data in compressed format for indexing and retrieval. A variation of dimensionality reduction which is commonly used for text data is known as latent semantic indexing. One of the interesting characteristics of latent semantic indexing is that it brings out the key semantic aspects of the text data, which makes it more suitable for a variety of mining applications. For example, the noise effects of synonymy and polysemy are reduced because of the use of such dimensionality reduction techniques. Another family of dimension reduction techniques are probabilistic topic models, notably PLSA, LDA, and their variants; they perform dimension reduction in a probabilistic way with potentially more meaningful topic representations based on word distributions. Supervised Learning Methods for Text Data: Supervised learning methods are general machine learning methods that can exploit training data (i.e., pairs of input data points and the corresponding desired output) to learn a classifier or regression function that can be used to compute predictions on unseen new data. Since a wide range of application problems can be cast as a classification problem (that can be solved using supervised learning), the problem of supervised learning is sometimes
also referred to as classification. Most of the traditional methods for text mining in the machine learning literature have been extended to solve problems of text mining. These include methods such as rule-based classifier, decision trees, nearest neighbor classifiers, maximum margin classifiers, and probabilistic classifiers.

Transfer Learning with Text Data: The afore-mentioned example of cross-lingual mining provides a case where the attributes of the text collection may be heterogeneous. Clearly, the feature representations in the different languages are heterogeneous, and it can often provide useful to transfer knowledge from one domain to another, especially when there is paucity of data in one domain. For example, labeled English documents are copious and easy to find. On the other hand, it is much harder to obtain labeled Chinese documents. The problem of transfer learning attempts to transfer the learned knowledge from one domain to another. Some other scenarios in which this arises is the case where we have a mixture of text and multimedia data. This is often the case in many web-based and social media applications such as Flickr, YouTube or other multimedia sharing sites. In such cases, it may be desirable to transfer the learned knowledge from one domain to another with the use of cross-media transfer.

Probabilistic Techniques for Text Mining: A variety of probabilistic methods, particularly unsupervised topic models such as PLSA and LDA and supervised learning methods such as conditional random fields are used frequently in the context of text mining algorithms. Since such methods are used frequently in a wide variety of contexts, it is useful to create an organized survey which describes the different tools and techniques that are used in this context.

Mining Text Streams: Many recent applications on the web create massive streams of text data. In particular web applications such as social networks which allow the simultaneous input of text from a wide variety of users can result in a continuous stream of large volumes of text data. Similarly, news streams such as Reuters or aggregators such as Google news create large volumes of streams which can be mined continuously. Such text data are more challenging to mine, because they need to be processed in the context of a one-pass constraint. The one-pass constraint essentially means that it may sometimes be difficult to store the data offline for processing, and it is necessary to perform the mining tasks continuously, as the data comes in. This makes algorithmic design a much more challenging task.

Cross-Lingual Mining of Text Data: With the proliferation of web-based and other information retrieval applications to other applications, it has become particularly useful to apply mining tasks in different languages or use the knowledge or corpora in one language to another. For example, in cross-language mining, it may be desirable to cluster a group of documents in different languages, so that documents from different languages but similar semantic topics may be placed in the same cluster. Such cross-lingual applications are extremely rich, because they can often be used to leverage knowledge from one data domain into another.

Text Mining in Multimedia Networks: Text often occurs in the context of many multimedia sharing sites such as Flickr or YouTube. A natural question arises as to whether we can enrich the underlying mining process by simultaneously using the data from other domains together with the text collection.

Text Mining in Social Media: One of the most common sources of text on the web is the presence of social media, which allows human actors to express themselves quickly and freely in the context of a wide range of subjects. Social media is now exploited widely by commercial sites for influencing users and targeted marketing. The process of mining text in social media requires the special ability to mine dynamic data which often contains poor and non-standard vocabulary. Furthermore, the text may occur in the context of linked social networks. Such links can be used in order to improve the quality of the underlying mining process.

Opinion Mining from Text Data: A considerable amount of text on web sites occurs in the context of product reviews or opinions of different users. Mining such opinionated text data to reveal and summarize the opinions about a topic has widespread applications, such as in supporting consumers for optimizing decisions and business intelligence. Spam opinions which are not useful and simply add noise to the mining process.

Text Mining from Biomedical Data: Text mining techniques play an important role in both enabling biomedical researchers to effectively and efficiently access the knowledge buried in large amounts of literature and supplementing the mining of other biomedical data such as genome sequences, gene expression data, and protein structures to facilitate and speed up biomedical discovery. As a result, a great deal of research work has been done in adapting and extending standard text mining methods to the biomedical domain, such as recognition of various biomedical entities and their relations, text summarization, and question answering.

4.2 A survey of Text Classification Algorithms

The problem of classification has been widely studied in the database, data mining, and information retrieval communities. The problem of classification is defined as follows. We have a set of training records *D* = *{X*1*, . . . , XN}*, such that each record is labeled with a class value drawn from a set of *k* different discrete values indexed by *{*1 *. . . k}*. The training data is used in order to construct a *classification model*, which relates the features in the underlying record to one of the class labels. For a given *test instance* for which the class is unknown, the training models used to predict a class label for this instance. In the *hard version* of the classification problem, a particular label is explicitly assigned to the instance, whereas in the *soft version* of the classification problem, a probability value is assigned to the test instance. Other variations of the classification problem allow ranking of different class choices for a test instance or allow the assignment of multiple labels to a test instance.

The classification problem assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set-valued features; however, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire lexicon size) is much greater than a typical set-valued classification problem. A number of the techniques discussed in this chapter have also been converted into software and are publicly available through multiple toolkits such as the BOW toolkit, Mallot, WEKA 1, and LingPipe 2.

The problem of text classification finds applications in a wide variety of domains in text mining. Some examples of domains in which text classification is commonly used are as follows:

News filtering and Organization: Most of the news services today are electronic in nature in which a large volume of news articles are created very single day by the organizations. In such cases, it is difficult to organize the news articles manually. Therefore, automated methods can be very useful for news categorization in a variety of web portals. This application is also referred to as *text filtering*.

Document Organization and Retrieval: The above application is generally useful for many applications beyond news filtering and organization. A variety of supervised methods may be used for document organization in many domains. These include large digital libraries of documents, web collections, scientific literature, or even social feeds. Hierarchically organized document collections can be particularly useful for browsing and retrieval.

Opinion Mining: Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review.

Email Classification and Spam Filtering: It is often desirable to classify email in order to determine either the subject or to determine junk email in an automated way. This is also referred to as *spam filtering* or *email filtering*. A wide variety of techniques have been designed for text classification. In this chapter, we will discuss the broad classes of techniques, and their uses for classification tasks. We note that these classes of techniques also generally exist for other data domains such as quantitative or categorical data. Since text may be modeled as quantitative data with frequencies on the word attributes, it is possible to use most of the methods for quantitative data directly on text. However, text is a particular kind of data in which the word attributes are sparse, and high dimensional, with low frequencies on most of the words. Therefore, it is critical to design classification methods which effectively account for these characteristics of text. In this chapter, we will focus on the specific changes which are applicable to the text domain. Some key methods, which are commonly used for text classification are as follows:

Decision Trees: Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features. The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution. For a given text instance, we determine the partition that it is most likely to belong to and use it for the purposes of classification.

Pattern (Rule)-based Classifiers: In rule-based classifiers we determine the word patterns which are most likely to be related to the different classes. We construct a set of rules, in which the left-hand side corresponds to a word pattern, and the right-hand side corresponds to a class label. These rules are used for the purposes of classification.

SVM Classifiers: SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification.

Neural Network Classifiers: Neural networks are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the *generative classifiers*.

Bayesian (Generative) Classifiers: In Bayesian classifiers (also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

Other Classifiers: Almost all classifiers can be adapted to the case of text data. Some of the other classifiers include nearest neighbor classifiers, and genetic algorithm-based classifiers. We will discuss some of these different classifiers in some detail and their use for the case of text data.

The area of text categorization is so vast that it is impossible to cover all the different algorithms in detail in a single chapter. Therefore, our goal is to provide the reader with an overview of the most important techniques, and also the pointers to the different variations of these techniques.

4.3 A survey of Text Clustering Algorithms

The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function. The problem of clustering can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms. Clustering is especially useful for organizing documents to improve retrieval and support browsing.

The study of the clustering problem precedes its applicability to the text domain. Traditional methods for clustering have generally focused on the case of quantitative data, in which the attributes of the data are numeric. The problem has also been studied for the case of categorical data, in which the attributes may take on nominal values.. A number of implementations of common text clustering algorithms, as applied to text data, may be found in several toolkits such as Lemur and BOW toolkit. The problem of clustering finds applicability for a number of tasks:

Document Organization and Browsing: The hierarchical organization of documents into coherent categories can be very useful for systematic browsing of the document collection. A classical example of this is the *Scatter/Gather* method, which provides a systematic browsing technique with the use of clustered organization of the document collection.

Corpus Summarization: Clustering techniques provide a coherent summary of the collection in the form of *cluster-digests* or *word-clusters*, which can be used in order to provide summary insights into the overall content of the underlying corpus. The problem of clustering is also closely related to that of dimensionality reduction and topic modeling. Such dimensionality reduction methods are all different ways of summarizing a corpus of documents.

Document Classification: While clustering is inherently an unsupervised learning method, it can be leveraged in order to improve the quality of the results in its supervised variant. In particular, word-clusters and co-training methods can be used in order to improve the classification accuracy of supervised applications with the use of clustering techniques.

We note that many classes of algorithms such as the k-means algorithm,

or hierarchical algorithms are general-purpose methods, which can be extended to any kind of data, including text data. A text document can be represented either in the form of binary data, when we use the presence or absence of a word in the document in order to create a binary vector. In such cases, it is possible to directly use a variety of categorical data clustering algorithms on the binary representation. A more enhanced representation would include refined weighting methods based on the frequencies of the individual words in the document as well as frequencies of words in an entire collection (e.g., TF-IDF weighting). Quantitative data clustering algorithms can be used in conjunction with these frequencies in order to determine the most relevant groups of objects in the data. However, such naive techniques do not typically work well for clustering text data. This is because text data has a number of unique properties which necessitate the design of specialized algorithms for the task. The distinguishing characteristics of the text representation are as follows: The dimensionality of the text representation is very large, but the underlying data is sparse. In other words, the lexicon from which the documents are drawn may be of the order of 105, but a given document may contain only a few hundred words. This problem is even more serious when the documents to be clustered are very short (e.g., when clustering sentences or tweets). While the lexicon of a given corpus of documents may be large, the words are typically correlated with one another. This means that the number of concepts (or principal components) in the data is much smaller than the feature space. This necessitates the careful design of algorithms which can account for word correlations in the clustering process. The number of words (or non-zero entries) in the different documents may vary widely. Therefore, it is important to normalize the document representations appropriately during the clustering task. The sparse and high dimensional representation of the different documents necessitate the design of text-specific algorithms for document representation and processing, a topic heavily studied in the information retrieval literature where many techniques have been proposed to optimize document representation for improving the accuracy of matching a document with a query. Most of these techniques can also be used to improve document representation for clustering.

In order to enable an effective clustering process, the word frequencies need to be normalized in terms of their relative frequency of presence in the document and over the entire collection. In general, a common representation used for text processing is the *vector-space based* TF-IDF representation. In the TF-IDF representation, the term frequency for each word is normalized by the *inverse document frequency*, or IDF. The inverse document frequency normalization reduces the weight of terms which occur more frequently in the collection. This reduces the importance of common terms in the collection, ensuring that the matching of documents be more influenced by that of more discriminative words which have relatively low frequencies in the collection. In addition, a sub-linear transformation function is often applied to the term frequencies in order to avoid the undesirable dominating effect of any single term that might be very frequent in a document. The work on document-normalization is itself a vast area of research

Text clustering algorithms are divided into a wide variety of different types such as agglomerative clustering algorithms, partitioning algorithms, and standard parametric modeling based methods such as the EM-algorithm. Furthermore, text representations may also be treated as strings (rather than bags of words). These different representations necessitate the design of different classes of clustering algorithms. Different clustering algorithms have different tradeoffs in terms of effectiveness and efficiency. In this chapter it's been discussed a wide variety of algorithms which are commonly used for text clustering. We will also discuss text clustering algorithms for related scenarios such as dynamic data, network-based text data and semi-supervised scenarios.

5 Smart Logistics: Applying data classification technologies in online consumer insights towards enhancing logistics and delivery processes

In chapter 3 we presented the requirements and we elicited from an extensive literature review. Here, we elaborate on the requirement 6 in order to convert it into specific pilot objective in order to provide a proof-of concept for it. The Objective that corresponds to requirement 6 is: "*To identify online consumers' problems and preferences, relevant with logistics procedure, extracting information by social media*". So, in this section, we describe a analysis and an application scenario that address to the above pilot objective and their respective results.

5.1 Objective

The concept of social media is top of the agenda for many business executives today. Decision makers, as well as consultants, try to identify ways in which firms can make profitable use of applications such as Wikipedia, YouTube, Facebook, Second Life, and Twitter. Yet despite this interest, there seems to be very limited understanding of what the term "Social Media" exactly means. Under the umbrella of the term social media, we put all the sites/networks in which we may read and write "user generated content", with online consumer reviews (OCRs) being a significant part of that content.

Although OCRs have helped consumers to know about the strengths and weaknesses of different products and find the ones that best suit their needs, they introduce a challenge for businesses to analyse them because of their volume, variety, velocity and veracity. The predictors of readership and helpfulness of OCR using a sentiment mining approach for big data analytics has been investigated only in the past few years; findings up to now indicate results that could be used as the basis for new business models and adjustments in several procedures followed by large enterprises, including the processes and models, used in the logistics field. Nevertheless, current methods used for sorting OCR may bias both their readership and helpfulness; thus, it is necessary to develop scalable automated systems for sorting and classification of big OCR data which will benefit both vendors and consumers.

The basic objective of this analysis is to

(1) study the importance of "user generated content" focusing on OCR,

(2) identify the basic KPIs with which firms may measure the content importance,

(3) develop the necessary system for sorting and classification of big OCR data and

(4) propose different ways with which e-commerce companies may use these KPIs and the aforementioned system, in order to extract useful information and adjust their logistics procedures.

To become more specific, we aim to develop a tool which will be able to fulfil the following sub-objectives:

▪ Understand consumers' behavior relevant to logistics process

▪ Understand the problematic parts of the online purchasing procedure

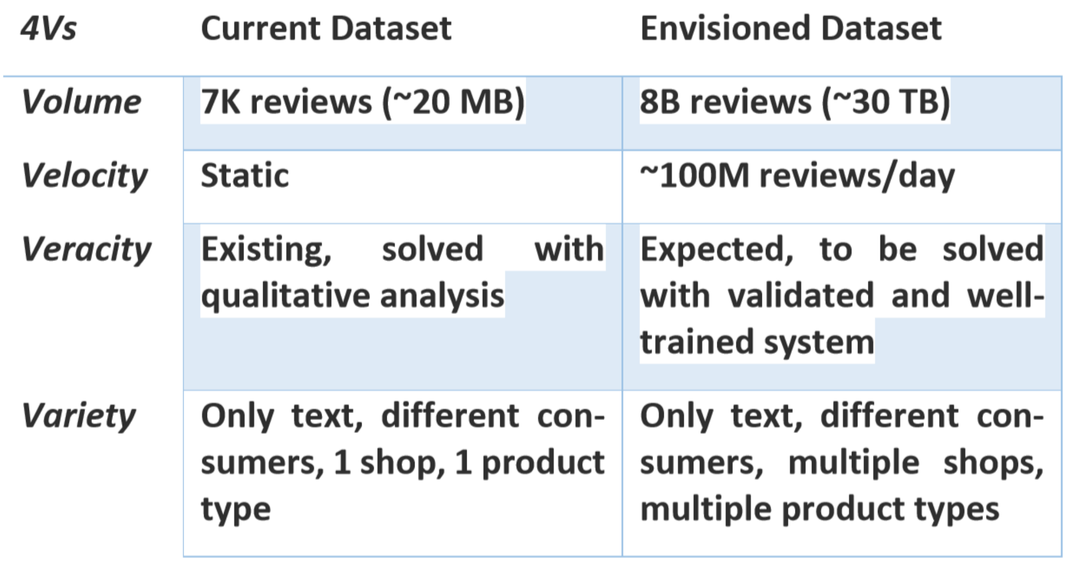
▪ Identify the main problems identified by customers

▪ Give to e-commerce firms the opportunity to ameliorate parts of its logistics procedure, as well as to develop new business models which will be able to minimize their "transportation"/ logistics cost

5.2 Methodology

The data comprise of text reviews from consumers; thus, it is what is called unstructured data. There is no model in structuring the opinions and reviews of consumers, expect from identifying the basic nature of the review: positive, negative or neutral.

The dataset used in the 1st step includes a pre-classification in positive and negative reviews. Regarding the 4Vs of Big Data, the current and envisioned datasets for the potential exploitation of this work in the future stand as described in Table 5.1**.**



4Vs	Current Dataset	Envisioned Dataset
Volume	**7K reviews (~20 MB)**	**8B reviews (~30 TB)**
Velocity	**Static**	**~100M reviews/day**
Veracity	**Existing, solved with qualitative analysis**	**Expected, to be solved with validated and well-trained system**
Variety	**Only text, different consumers, 1 shop, 1 product type**	**Only text, different consumers, multiple shops, multiple product types**

Table 5.1 - *Current and envisioned dataset in regard to the 4Vs*

In order to justify the details of the envisioned dataset, let us present the case of one example. Amazon, published the 2nd Quarter 2016 statistics, according to which it comprises:

▪ 244 Million active buyer accounts (54 Million are Prime buyers) - up 20% in a year

▪ 200 Million active products on Amazon

▪ 2.2 Billion sales in the past 12 months (average 6 Million sales a day)

o Non-peak holiday sales per day: 4.3 Million
o ~ 50 sales per second

Amazon provides access to the consumers' reviews. Only in Amazon, for items like electronics and computers and accessories, the average product has more than 4,000 reviews.

Regarding Veracity, consumer reviews may be misleading: the tool will be trained to identify real problems in reviews. Finally, in regards to variety, the consumers' reviews comprise only text, but from different sources. In unstructured data like reviews of consumers each consumer may be defined as a separate source. Definitely different types of e-shops will be used as well, along with different types of products. In the current dataset, only electronics were investigated, from one particular e-shop in Greece.

The methodology which is followed to apply analytics tools in this use case is separated into three steps (Figure 5.2):

▪ The identification of the problems, using text mining and meaning mining procedures
▪ The training of the text mining tool, in order to be able to accurately identify the problems in large data sets
▪ The application of the tool in larger data sets

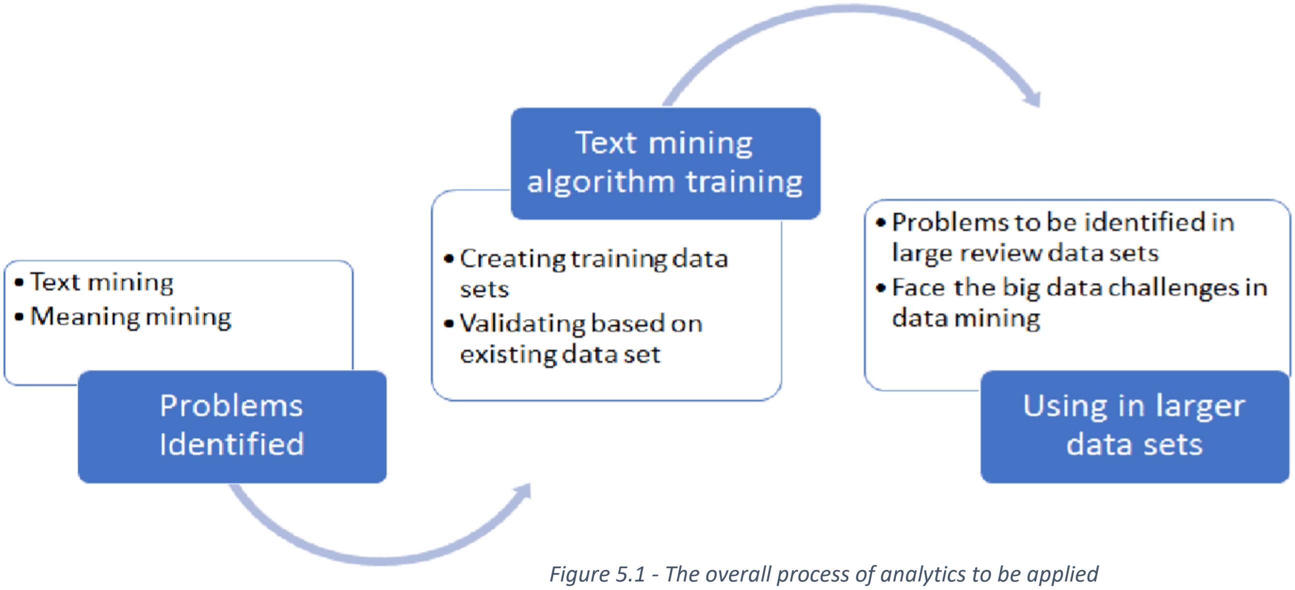

• Text mining
• Meaning mining
Problems Identified
Text mining algorithm training
• Creating training data sets
• Validating based on existing data set
• Problems to be identified in large review data sets
• Face the big data challenges in data mining
Using in larger data sets

Figure 5.1 - The overall process of analytics to be applied

5.3 Results

- 1st step and results up to now

Based on the current e-commerce environment we choose the sectors that consist high priority for the e-commerce economy and at the same time confront the major problems in logistic sector. The first sector that has been chosen is electronics.

A crawler is developed in order to extract consumers' reviews by one of the biggest price comparison sites. We collect public reviews for technology e-shops. The crawler parses user reviews and stores the data in preformatted documents with attributes such as: shop name, shop total votes, date of the review, positive or negative key features (selected form predefined lists), review main text, star rating.

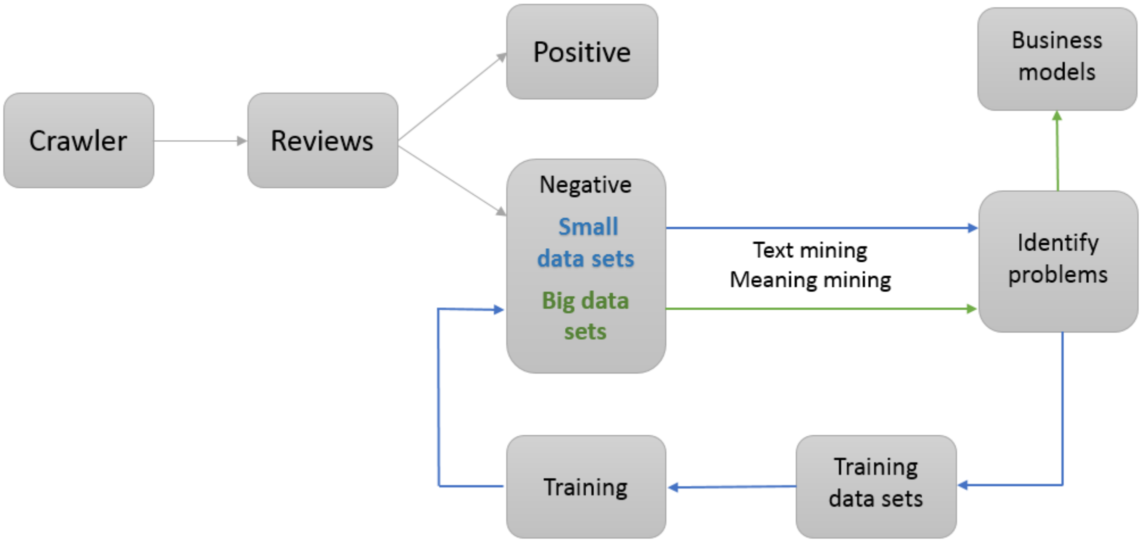
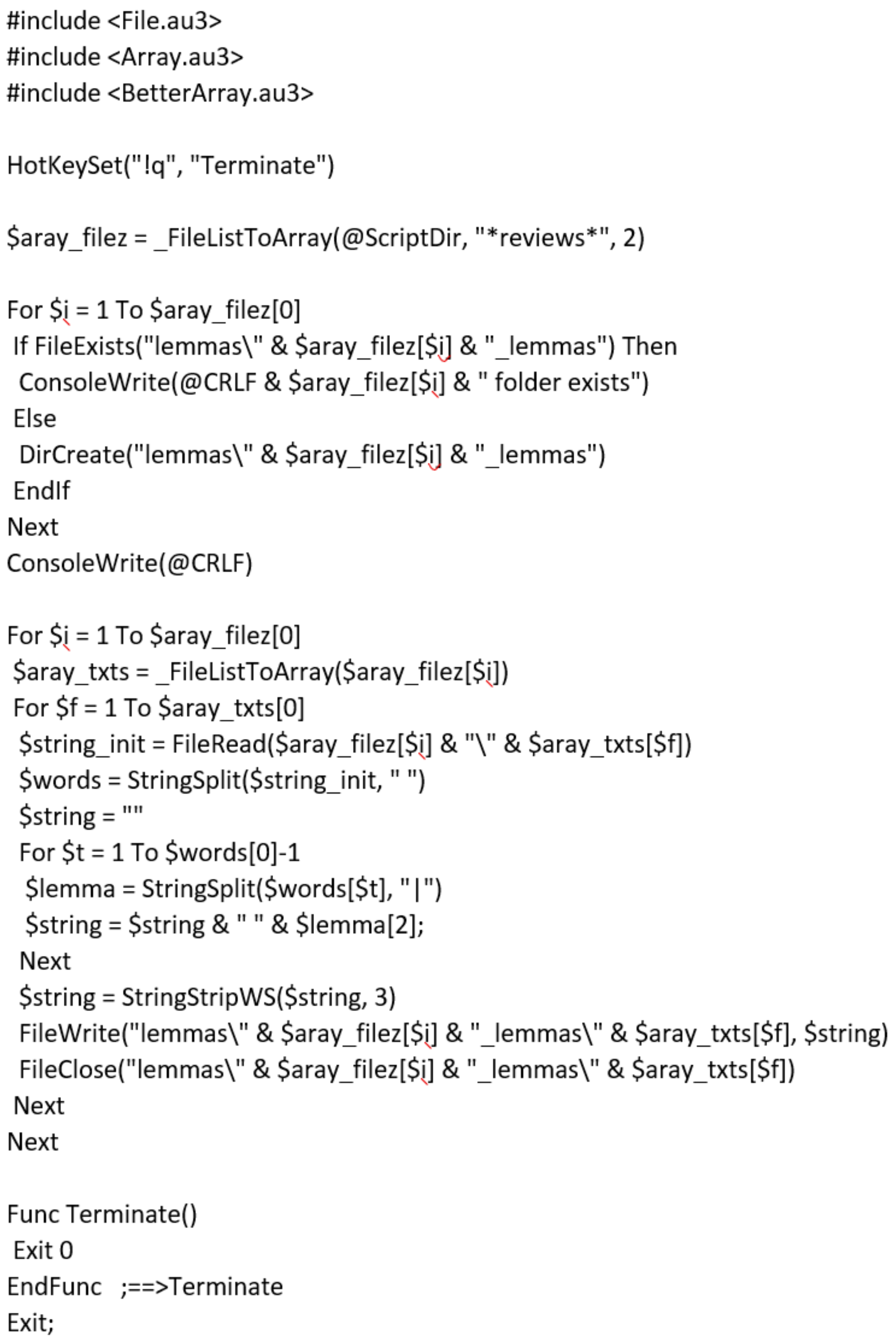


Figure 5.2 – Process flow chart

In the first part of the analysis procedure, we separate negative and positive reviews and the most used words in these posts have been identified. We convert words in each review to lemmas (example: "I loved customer service" is transformed to "Be love customer service". We also perform part-of-speech tagging and keep only nouns, adjectives and adverbs with the sentiment (consumer insights), and group user reviews by the occurrence of these insights.

An algorithm has been created which converts words in lemmas by using Natural Language Processing services developed by the NLP group of the Institute for Language and Speech Processing enriched with Greek word library and Apache OpenNLP. Apache OpenNLP is an open source Java library which is used process Natural Language text. OpenNLP provides services such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution, etc.



```
#include <File.au3>
#include <Array.au3>
#include <BetterArray.au3>

HotKeySet("!q", "Terminate")

$aray_filez = _FileListToArray(@ScriptDir, "*reviews*", 2)

For $i = 1 To $aray_filez[0]
 If FileExists("lemmas\" & $aray_filez[$i] & "_lemmas") Then
  ConsoleWrite(@CRLF & $aray_filez[$i] & " folder exists")
 Else
  DirCreate("lemmas\" & $aray_filez[$i] & "_lemmas")
 EndIf
Next
ConsoleWrite(@CRLF)

For $i = 1 To $aray_filez[0]
 $aray_txts = _FileListToArray($aray_filez[$i])
 For $f = 1 To $aray_txts[0]
  $string_init = FileRead($aray_filez[$i] & "\" & $aray_txts[$f])
  $words = StringSplit($string_init, " ")
  $string = ""
  For $t = 1 To $words[0]-1
   $lemma = StringSplit($words[$t], "|")
   $string = $string & " " & $lemma[2];
  Next
  $string = StringStripWS($string, 3)
  FileWrite("lemmas\" & $aray_filez[$i] & "_lemmas\" & $aray_txts[$f], $string)
  FileClose("lemmas\" & $aray_filez[$i] & "_lemmas\" & $aray_txts[$f])
 Next
Next

Func Terminate()
 Exit 0
EndFunc   ;==>Terminate
Exit;
```

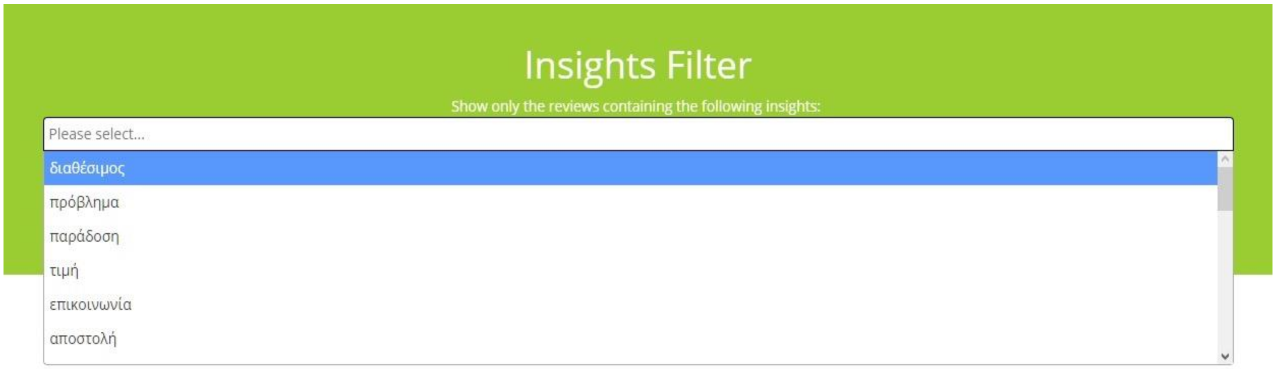
Reviews before and after the Lemmatizer.

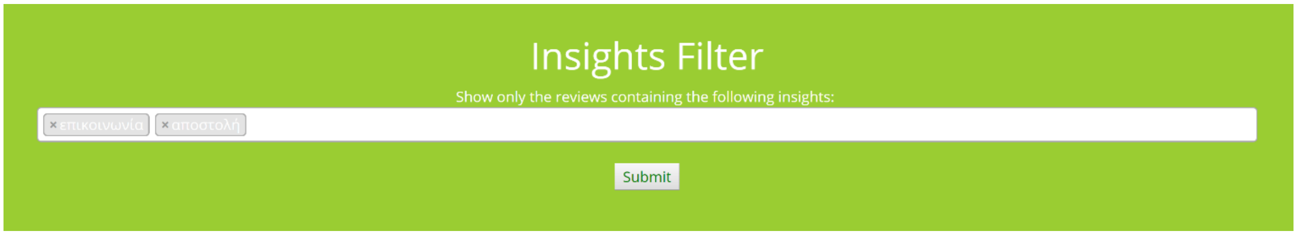
```
Παρήγγειλα σκούπα Swivel G6, ήρθε Swivel max με
χαλασμένο φορτιστή. Την επέστρεψα την επομένη με δική
τους υπόδειξη για έλεγχο, και μου επεστράφει μετά από
κάποιες μέρες η ίδια σκούπα γιατί κατά τη γνώμη τους
είχε χρησιμοποιηθεί με άλλο(;) φορτιστή και τη μπαταρία
ξεκολημένη...
```

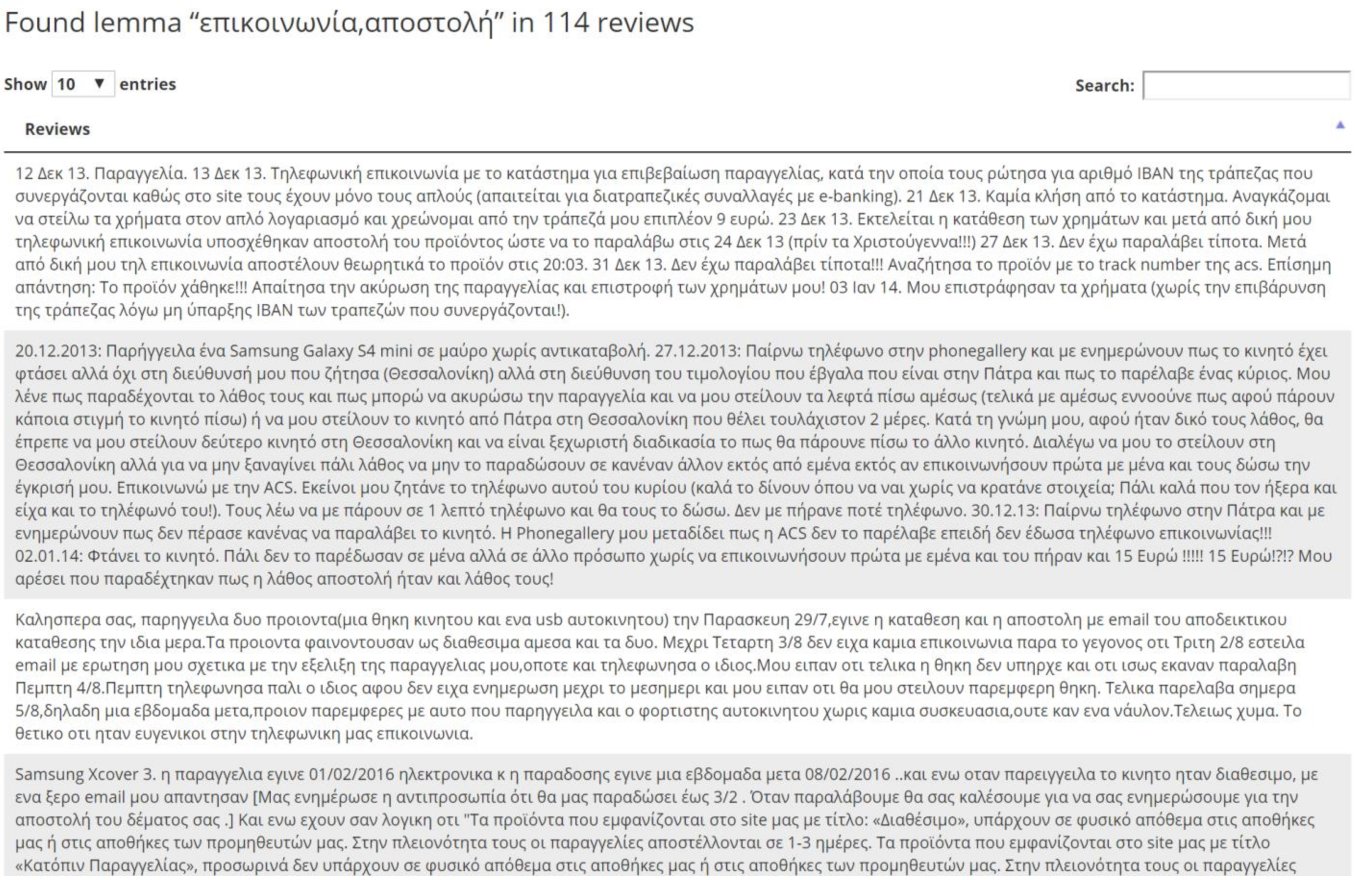
```
Παρήγγειλα|παραγγέλλω|VbMnIdPa01SgXxPeAvXx
σκούπα|σκούπα|NoCmFeSgAc Swivel|Swivel|RgFwOr G6|G6
|RgFwOr ,|,|PUNCT ήρθε|έρχομαι|VbMnIdPa03SgXxPePvXx
Swivel|Swivel|RgFwOr max|max|RgFwOr με|με|AsPpSp
χαλασμένο|χαλασμένος|AjBaMaSgAc
φορτιστή|φορτιστής|NoCmMaSgAc
επέστρεψα|επιστρέφω|VbMnIdPa01SgXxPeAvXx
την|ο|AtDfFeSgAc επομένη|επόμενος|AjBaFeSgAc
με|με|AsPpSp δική|δικός|AjBaFeSgAc
τους|μου|PnPoMa03PlGeXx υπόδειξη|υπόδειξη|NoCmFeSgAc
για|για|AsPpSp έλεγχο|έλεγχος|NoCmMaSgAc ,|,|PUNCT
και|και|CjCo μου|εγώ|PnPeMa01SgGeWe
επεστράφει|επιστρέφω|VbMnIdPr03SgXxIpAvXx
μετά|μετά|AdXxBa από|από|AsPpSp
κάποιες|κάποιος|PnIdFe03PlAcXx μέρες|μέρα|NoCmFePlAc
η|ο|AtDfFeSgNm ίδια|ίδιος|AjBaFeSgNm
σκούπα|σκούπα|NoCmFeSgNm γιατί|γιατί|CjSb
κατά|κατά|AsPpSp τη|ο|AtDfFeSgAc γνώμη|γνώμη|NoCmFeSgAc
τους|μου|PnPoMa03PlGeXx είχε|έχω|VbMnIdPa03SgXxIpAvXx
χρησιμοποιηθεί|χρησιμοποιώ|VbMnNfXxXxXxXxPePvXx
με|με|AsPpSp άλλο|άλλος|PnIdNe03SgAcXx (|(|OPUNCT ;|;
|PTERM_P )|)|CPUNCT φορτιστή|φορτιστής|NoCmMaSgAc
και|και|CjCo τη|ο|AtDfFeSgAc
μπαταρία|μπαταρία|NoCmFeSgAc
ξεκολημένη|ξεκολημένη|VbMnPpXxXxSgFePePvAc ...|...
|PTERM_P
```

```
παραγγέλλω σκούπα Swivel G6 , έρχομαι Swivel max με χαλασμένος
φορτιστής επιστρέφω ο επόμενος με δικός μου υπόδειξη για έλεγχος ,
και εγώ επιστρέφω μετά από κάποιος μέρα ο ίδιος σκούπα γιατί κατά ο
γνώμη μου έχω χρησιμοποιώ με άλλος ( ; ) φορτιστής και ο μπαταρία
ξεκολημένη ...
```

In the second part of the analysis procedure, the previous words are combined (via text mining tools) in order to understand better the way that these words may explain consumer behavior.







Found lemma “επικοινωνία,αποστολή” in 114 reviews

Reviews
12 Δεκ 13. Παραγγελία. 13 Δεκ 13. Τηλεφωνική επικοινωνία με το κατάστημα για επιβεβαίωση παραγγελίας, κατά την οποία τους ρώτησα για αριθμό IBAN της τράπεζας που συνεργάζονται καθώς στο site τους έχουν μόνο τους απλούς (απαιτείται για διατραπεζικές συναλλαγές με e-banking). 21 Δεκ 13. Καμία κλήση από το κατάστημα. Αναγκάζομαι να στείλω τα χρήματα στον απλό λογαριασμό και χρεώνομαι από την τράπεζά μου επιπλέον 9 ευρώ. 23 Δεκ 13. Εκτελείται η κατάθεση των χρημάτων και μετά από δική μου τηλεφωνική επικοινωνία υποσχέθηκαν αποστολή του προϊόντος ώστε να το παραλάβω στις 24 Δεκ 13 (πρίν τα Χριστούγεννα!!!) 27 Δεκ 13. Δεν έχω παραλάβει τίποτα. Μετά από δική μου τηλ επικοινωνία αποστέλουν θεωρητικά το προϊόν στις 20:03. 31 Δεκ 13. Δεν έχω παραλάβει τίποτα!!! Αναζήτησα το προϊόν με το track number της acs. Επίσημη απάντηση: Το προϊόν χάθηκε!!! Απαίτησα την ακύρωση της παραγγελίας και επιστροφή των χρημάτων μου! 03 Ιαν 14. Μου επιστράφησαν τα χρήματα (χωρίς την επιβάρυνση της τράπεζας λόγω μη ύπαρξης IBAN των τραπεζών που συνεργάζονται!).
20.12.2013: Παρήγγειλα ένα Samsung Galaxy S4 mini σε μαύρο χωρίς αντικαταβολή. 27.12.2013: Παίρνω τηλέφωνο στην phonegallery και με ενημερώνουν πως το κινητό έχει φτάσει αλλά όχι στη διεύθυνσή μου που ζήτησα (Θεσσαλονίκη) αλλά στη διεύθυνση του τιμολογίου που έβγαλα που είναι στην Πάτρα και πως το παρέλαβε ένας κύριος. Μου λένε πως παραδέχονται το λάθος τους και πως μπορώ να ακυρώσω την παραγγελία και να μου στείλουν τα λεφτά πίσω αμέσως (τελικά με αμέσως εννοούνε πως αφού πάρουν κάποια στιγμή το κινητό πίσω) ή να μου στείλουν το κινητό από Πάτρα στη Θεσσαλονίκη που θέλει τουλάχιστον 2 μέρες. Κατά τη γνώμη μου, αφού ήταν δικό τους λάθος, θα έπρεπε να μου στείλουν δεύτερο κινητό στη Θεσσαλονίκη και να είναι ξεχωριστή διαδικασία το πως θα πάρουνε πίσω το άλλο κινητό. Διαλέγω να μου το στείλουν στη Θεσσαλονίκη αλλά για να μην ξαναγίνει πάλι λάθος να μην το παραδώσουν σε κανέναν άλλον εκτός από εμένα εκτός αν επικοινωνήσουν πρώτα με μένα και τους δώσω την έγκρισή μου. Επικοινωνώ με την ACS. Εκείνοι μου ζητάνε το τηλέφωνο αυτού του κυρίου (καλά το δίνουν όπου να ναι χωρίς να κρατάνε στοιχεία; Πάλι καλά που τον ήξερα και είχα και το τηλέφωνό του!). Τους λέω να με πάρουν σε 1 λεπτό τηλέφωνο και θα τους το δώσω. Δεν με πήρανε ποτέ τηλέφωνο. 30.12.13: Παίρνω τηλέφωνο στην Πάτρα και με ενημερώνουν πως δεν πέρασε κανένας να παραλάβει το κινητό. Η Phonegallery μου μεταδίδει πως η ACS δεν το παρέλαβε επειδή δεν έδωσα τηλέφωνο επικοινωνίας!!! 02.01.14: Φτάνει το κινητό. Πάλι δεν το παρέδωσαν σε μένα αλλά σε άλλο πρόσωπο χωρίς να επικοινωνήσουν πρώτα με εμένα και του πήραν και 15 Ευρώ !!!!! 15 Ευρώ!?!? Μου αρέσει που παραδέχτηκαν πως η λάθος αποστολή ήταν και λάθος τους!
Καλησπερα σας, παρηγγειλα δυο προιοντα(μια θηκη κινητου και ενα usb αυτοκινητου) την Παρασκευη 29/7,εγινε η καταθεση και η αποστολη με email του αποδεικτικου καταθεσης την ιδια μερα.Τα προιοντα φαινοντουσαν ως διαθεσιμα αμεσα και τα δυο. Μεχρι Τεταρτη 3/8 δεν ειχα καμια επικοινωνια παρα το γεγονος οτι Τριτη 2/8 εστειλα email με ερωτηση μου σχετικα με την εξελιξη της παραγγελιας μου,οποτε και τηλεφωνησα ο ιδιος.Μου ειπαν οτι τελικα η θηκη δεν υπηρχε και οτι ισως εκαναν παραλαβη Πεμπτη 4/8.Πεμπτη τηλεφωνησα παλι ο ιδιος αφου δεν ειχα ενημερωση μεχρι το μεσημερι και μου ειπαν οτι θα μου στειλουν παρεμφερη θηκη. Τελικα παρελαβα σημερα 5/8,δηλαδη μια εβδομαδα μετα,προιον παρεμφερες με αυτο που παρηγγειλα και ο φορτιστης αυτοκινητου χωρις καμια συσκευασια,ουτε καν ενα νάυλον.Τελειως χυμα. Το θετικο οτι ηταν ευγενικοι στην τηλεφωνικη μας επικοινωνια.
Samsung Xcover 3. η παραγγελια εγινε 01/02/2016 ηλεκτρονικα κ η παραδοσης εγινε μια εβδομαδα μετα 08/02/2016 ..και ενω οταν παρειγγειλα το κινητο ηταν διαθεσιμο, με ενα ξερο email μου απαντησαν [Μας ενημέρωσε η αντιπροσωπία ότι θα μας παραδώσει έως 3/2 . Όταν παραλάβουμε θα σας καλέσουμε για να σας ενημερώσουμε για την αποστολή του δέματος σας .] Και ενω εχουν σαν λογικη οτι "Τα προϊόντα που εμφανίζονται στο site μας με τίτλο: «Διαθέσιμο», υπάρχουν σε φυσικό απόθεμα στις αποθήκες μας ή στις αποθήκες των προμηθευτών μας. Στην πλειονότητα τους οι παραγγελίες αποστέλλονται σε 1-3 ημέρες. Τα προϊόντα που εμφανίζονται στο site μας με τίτλο «Κατόπιν Παραγγελίας», προσωρινά δεν υπάρχουν σε φυσικό απόθεμα στις αποθήκες μας ή στις αποθήκες των προμηθευτών μας. Στην πλειονότητα τους οι παραγγελίες

Text mining refers to

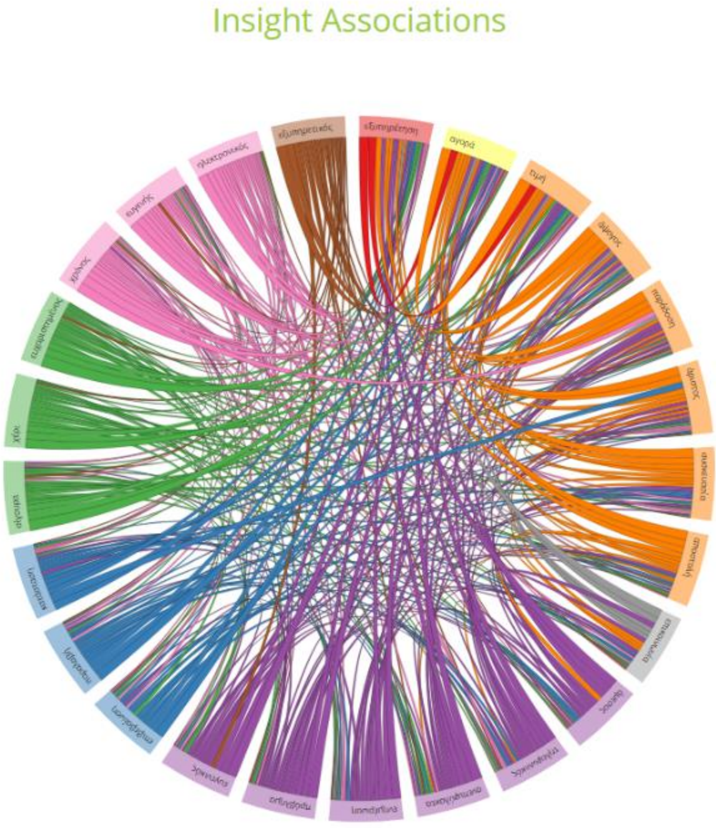
(i) identifying the frequencies of specific, common words that refer to specific problems identified in logistics procedures by consumers

(ii) quantifying the associations between the most important consumer insights, by providing correlations between the combination of specific keywords.

Associations are also visualised with an interactive 'cycle of associations':

Hovering the cursor over the words of the circle, we can see the correlations between the words and how frequently they appear together in customers' comments.

The thickness of the connection line/curve shows the intensity of the appearance in the consumers' reviews. In this way, the tool could be used to facilitate the problem identification based on the keywords' combination.



A deeper analysis conducted (meaning analysis) in order to identify the level of the importance of the problems that have been raised.

The aforementioned procedure is depicted in Figure 5.3, and it captures the current state of the described use case.

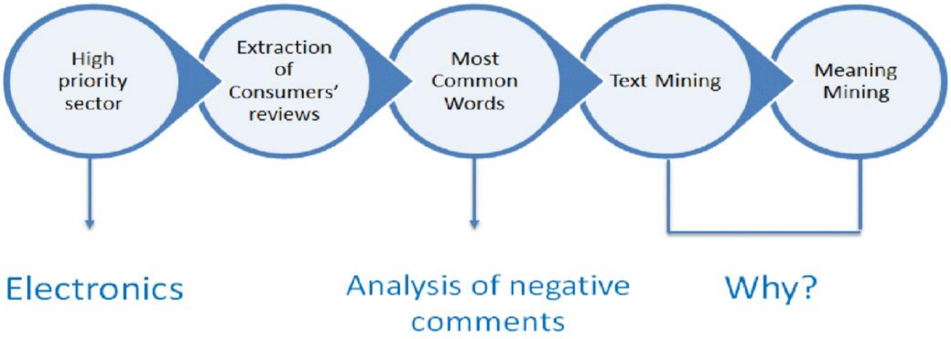


Figure 5.3 - Analysis Procedure

This procedure has already provided some first outcomes, that have been used to identify the steps to follow. These results are summarized in the following:

<u>Identification of the most common words and main problems</u>

Starting the analysis of users' comments, we identify which are the most common words and at the same time what is the correlation between them (Figure 5.4).

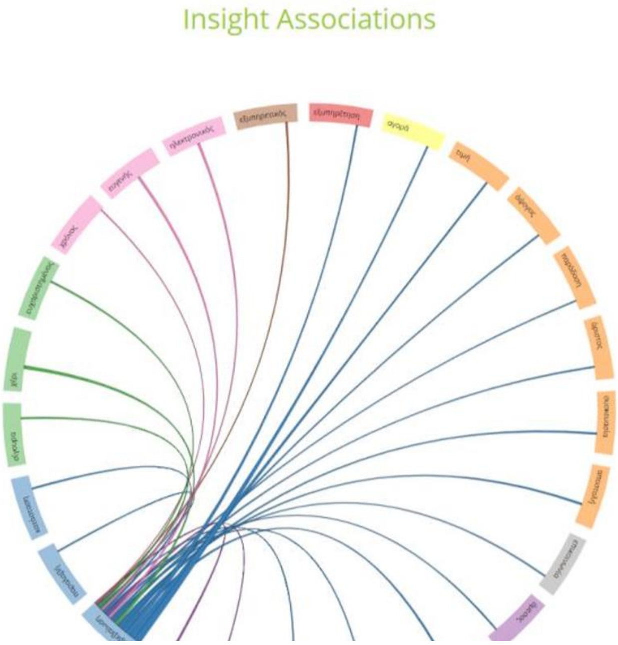


Figure 5.4 - Correlation between the most common words

This part of the analysis helps us to design the first picture of the problematic parts of logistics process.

Available	Problem	Delivery	Price
Communication	Purchase	Money	e-mail
Response	Cost	Card	Transport
Packaging	Time	Procedure	Date

Table 5.2 - Most Common words

Trying to interpret the meaning of the previous analysis, we may identify three main consumers' concerns relevant to logistics:

- Product availability
- Communication Issues
- Cost and payment issues

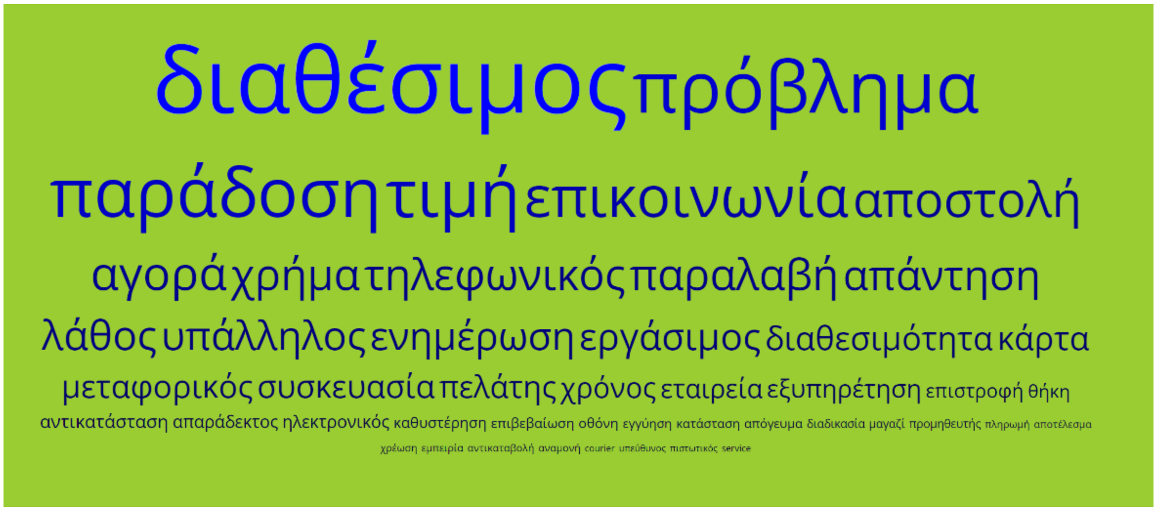

διαθέσιμος πρόβλημα
παράδοση τιμή επικοινωνία αποστολή
αγορά χρήμα τηλεφωνικός παραλαβή απάντηση
λάθος υπάλληλος ενημέρωση εργάσιμος διαθεσιμότητα κάρτα
μεταφορικός συσκευασία πελάτης χρόνος εταιρεία εξυπηρέτηση επιστροφή θήκη
αντικατάσταση απαράδεκτος ηλεκτρονικός καθυστέρηση επιβεβαίωση οθόνη εγγύηση κατάσταση απόγευμα διαδικασία μαγαζί προμηθευτής πληρωμή αποτέλεσμα
χρέωση εμπειρία αντικαταβολή αναμονή courier υπεύθυνος πιστωτικός service

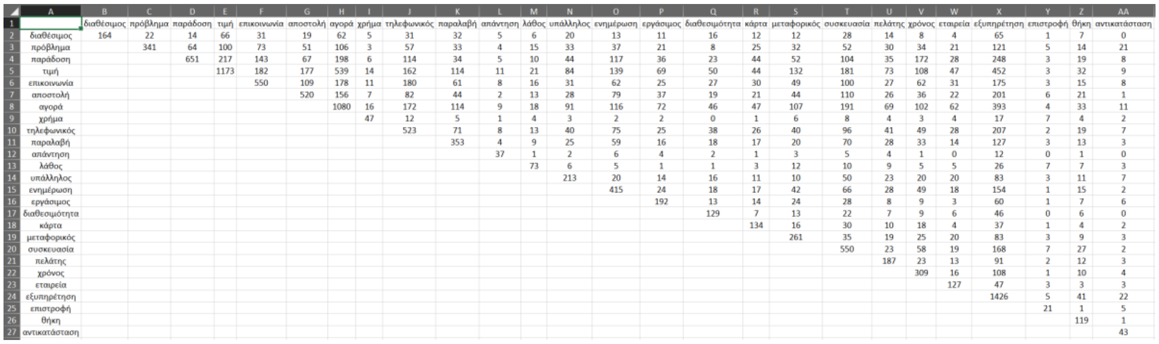
Figure 5.5 - Most Common words

Comparing this analysis with the results that have been extracted by the survey that has been conducted by ELTRUN (2016), we may point out that these are the basic issues that e-shops should be confronted with.

Text mining and meaning mining

Having identified the basic issues, the next question that has to be answered is relevant with the importance of each issue. Two different techniques are used:

1. Correlation between common words



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1		διαθέσιμος	πρόβλημα	παράδοση	τιμή	επικοινωνία	αποστολή	αγορά	χρήμα	τηλεφωνικός	παραλαβή	απάντηση	λάθος	υπάλληλος	ενημέρωση	εργάσιμος	διαθεσιμότητα	κάρτα	μεταφορικός	συσκευασία	πελάτης	χρόνος	εταιρεία	εξυπηρέτηση	επιστροφή	θήκη	αντικατάσταση
2	διαθέσιμος	164	22	14	66	31	19	62	5	31	32	5	6	20	13	11	16	12	12	28	14	8	4	65	1	7	0
3	πρόβλημα		341	64	100	73	51	106	3	57	33	4	15	33	37	21	8	25	32	52	30	34	21	121	5	14	21
4	παράδοση			651	217	143	67	198	6	114	34	5	10	44	117	36	23	44	52	104	35	172	28	248	3	19	8
5	τιμή				1173	182	177	539	14	162	114	11	21	84	139	69	50	44	132	181	73	108	47	452	3	32	9
6	επικοινωνία					550	109	178	11	180	61	8	16	31	62	25	27	30	49	100	27	62	31	175	3	15	8
7	αποστολή						520	156	7	82	44	2	13	28	79	37	19	21	44	110	26	36	22	201	6	21	1
8	αγορά							1080	16	172	114	9	18	91	116	72	46	47	107	191	69	102	62	393	4	33	11
9	χρήμα								47	12	5	1	4	3	2	2	0	1	6	8	4	3	4	17	7	4	2
10	τηλεφωνικός									523	71	8	13	40	75	25	38	26	40	96	41	49	28	207	2	19	7
11	παραλαβή										353	4	9	25	59	16	18	17	20	70	28	33	14	127	3	13	3
12	απάντηση											37	1	2	6	4	2	1	3	5	4	1	0	12	0	1	0
13	λάθος												73	6	5	1	1	3	12	10	9	5	5	26	7	7	3
14	υπάλληλος													213	20	14	16	11	10	50	23	20	20	83	3	11	7
15	ενημέρωση														415	24	18	17	42	66	28	49	18	154	1	15	2
16	εργάσιμος															192	13	14	24	28	8	9	3	60	1	7	6
17	διαθεσιμότητα																129	7	13	22	7	9	6	46	0	6	0
18	κάρτα																	134	16	30	10	18	4	37	1	4	2
19	μεταφορικός																		261	35	19	25	20	83	3	9	3
20	συσκευασία																			550	23	58	19	168	7	27	2
21	πελάτης																				187	23	13	91	2	12	3
22	χρόνος																					309	16	108	1	10	4
23	εταιρεία																						127	47	3	3	3
24	εξυπηρέτηση																							1426	5	41	22
25	επιστροφή																								21	1	5
26	θήκη																									119	1
27	αντικατάσταση																										43

2. Meaning mining technique (analysis based on the meaning/ logic of every comment)

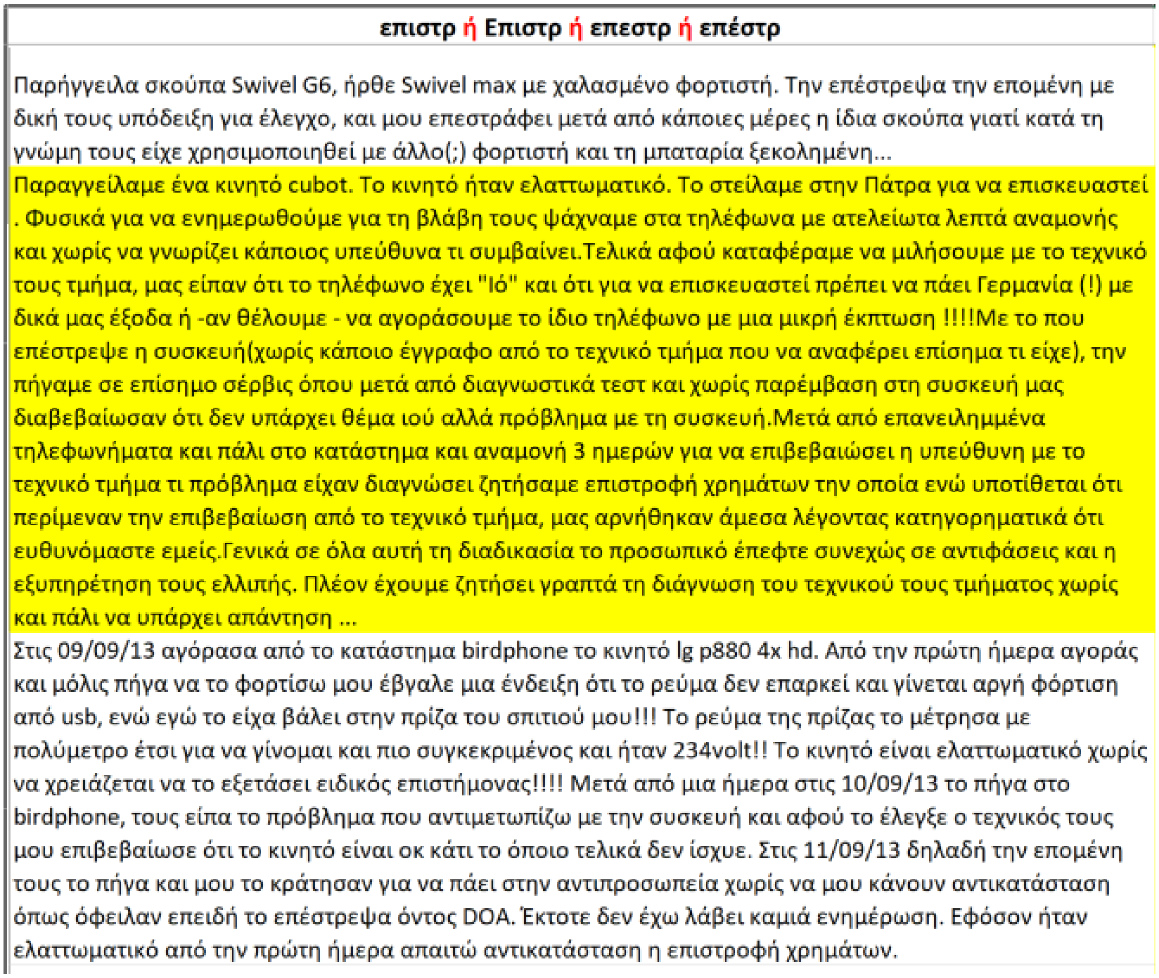
The main result of this part of the analysis, comes from the difference that is noticed between the two techniques. Table 8 describes an example of the results:

Main word	Association	Correlation	Meaning Mining
Return	Money	12,7%	21%
Return	Purchase	9,5%	14%
Return	Communication	5,9%	11%
Return	**Receive (product)**	**3,8%**	**25%**

Table 5.3 - Analysis Results

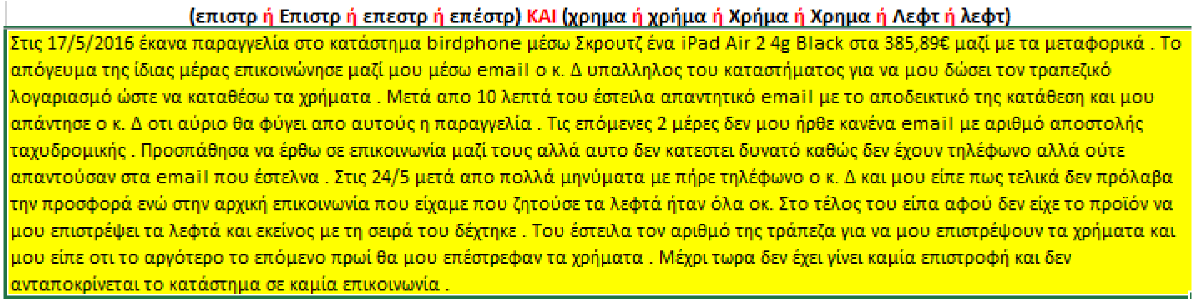
As we may notice, a difference exists in the results that were extracted based on the two techniques. This difference is justified; the choice of keywords, the possibility of users using synonyms and several other aspects may lead to the text mining outcomes vary somehow from the actual meaning mining outcomes. In other words, the problems actually identified via the qualitative analysis and the meaning mining process are not always captured by the basic text analysis tools. Typical example: The combination of the words “return” and “money” refer to a specific problem, having to do with problematic procedures in terms of refunds when a product is returned.

Text mining has identified that these two words' correlation is 12.7%, meaning that the word 'money' can be found in 12.7% of the reviews that refer to returns.



επιστρ ή Επιστρ ή επεστρ ή επέστρ
Παρήγγειλα σκούπα Swivel G6, ήρθε Swivel max με χαλασμένο φορτιστή. Την επέστρεψα την επομένη με δική τους υπόδειξη για έλεγχο, και μου επεστράφει μετά από κάποιες μέρες η ίδια σκούπα γιατί κατά τη γνώμη τους είχε χρησιμοποιηθεί με άλλο(;) φορτιστή και τη μπαταρία ξεκολημένη...
Παραγγείλαμε ένα κινητό cubot. Το κινητό ήταν ελαττωματικό. Το στείλαμε στην Πάτρα για να επισκευαστεί . Φυσικά για να ενημερωθούμε για τη βλάβη τους ψάχναμε στα τηλέφωνα με ατελείωτα λεπτά αναμονής και χωρίς να γνωρίζει κάποιος υπεύθυνα τι συμβαίνει.Τελικά αφού καταφέραμε να μιλήσουμε με το τεχνικό τους τμήμα, μας είπαν ότι το τηλέφωνο έχει "Ιό" και ότι για να επισκευαστεί πρέπει να πάει Γερμανία (!) με δικά μας έξοδα ή -αν θέλουμε - να αγοράσουμε το ίδιο τηλέφωνο με μια μικρή έκπτωση !!!!Με το που επέστρεψε η συσκευή(χωρίς κάποιο έγγραφο από το τεχνικό τμήμα που να αναφέρει επίσημα τι είχε), την πήγαμε σε επίσημο σέρβις όπου μετά από διαγνωστικά τεστ και χωρίς παρέμβαση στη συσκευή μας διαβεβαίωσαν ότι δεν υπάρχει θέμα ιού αλλά πρόβλημα με τη συσκευή.Μετά από επανειλημμένα τηλεφωνήματα και πάλι στο κατάστημα και αναμονή 3 ημερών για να επιβεβαιώσει η υπεύθυνη με το τεχνικό τμήμα τι πρόβλημα είχαν διαγνώσει ζητήσαμε επιστροφή χρημάτων την οποία ενώ υποτίθεται ότι περίμεναν την επιβεβαίωση από το τεχνικό τμήμα, μας αρνήθηκαν άμεσα λέγοντας κατηγορηματικά ότι ευθυνόμαστε εμείς.Γενικά σε όλα αυτή τη διαδικασία το προσωπικό έπεφτε συνεχώς σε αντιφάσεις και η εξυπηρέτηση τους ελλιπής. Πλέον έχουμε ζητήσει γραπτά τη διάγνωση του τεχνικού τους τμήματος χωρίς και πάλι να υπάρχει απάντηση ...
Στις 09/09/13 αγόρασα από το κατάστημα birdphone το κινητό lg p880 4x hd. Από την πρώτη ημέρα αγοράς και μόλις πήγα να το φορτίσω μου έβγαλε μια ένδειξη ότι το ρεύμα δεν επαρκεί και γίνεται αργή φόρτιση από usb, ενώ εγώ το είχα βάλει στην πρίζα του σπιτιού μου!!! Το ρεύμα της πρίζας το μέτρησα με πολύμετρο έτσι για να γίνομαι και πιο συγκεκριμένος και ήταν 234volt!! Το κινητό είναι ελαττωματικό χωρίς να χρειάζεται να το εξετάσει ειδικός επιστήμονας!!!! Μετά από μια ημέρα στις 10/09/13 το πήγα στο birdphone, τους είπα το πρόβλημα που αντιμετωπίζω με την συσκευή και αφού το έλεγξε ο τεχνικός τους μου επιβεβαίωσε ότι το κινητό είναι οκ κάτι το οποίο τελικά δεν ίσχυε. Στις 11/09/13 δηλαδή την επομένη τους το πήγα και μου το κράτησαν για να πάει στην αντιπροσωπεία χωρίς να μου κάνουν αντικατάσταση όπως όφειλαν επειδή το επέστρεψα όντος DOA. Έκτοτε δεν έχω λάβει καμιά ενημέρωση. Εφόσον ήταν ελαττωματικό από την πρώτη ημέρα απαιτώ αντικατάσταση η επιστροφή χρημάτων.

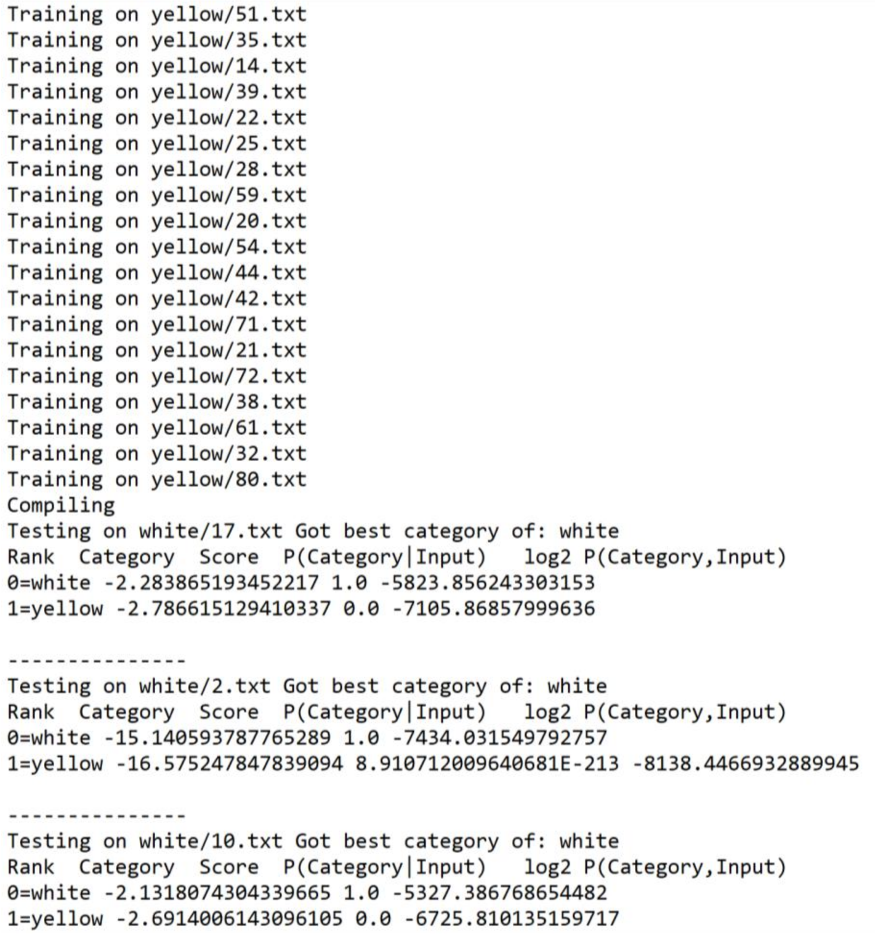
On the other hand, the qualitative analysis and the meaning mining process revealed that from the reviews that refer to returns, 21% actually mention that specific problem.



(επιστρ ή Επιστρ ή επεστρ ή επέστρ) ΚΑΙ (χρημα ή χρήμα ή Χρήμα ή Χρημα ή Λεφτ ή λεφτ)
Στις 17/5/2016 έκανα παραγγελία στο κατάστημα birdphone μέσω Σκρουτζ ένα iPad Air 2 4g Black στα 385,89€ μαζί με τα μεταφορικά . Το απόγευμα της ίδιας μέρας επικοινώνησε μαζί μου μέσω email ο κ. Δ υπαλληλος του καταστήματος για να μου δώσει τον τραπεζικό λογαριασμό ώστε να καταθέσω τα χρήματα . Μετά απο 10 λεπτά του έστειλα απαντητικό email με το αποδεικτικό της κατάθεση και μου απάντησε ο κ. Δ οτι αύριο θα φύγει απο αυτούς η παραγγελία . Τις επόμενες 2 μέρες δεν μου ήρθε κανένα email με αριθμό αποστολής ταχυδρομικής . Προσπάθησα να έρθω σε επικοινωνία μαζί τους αλλά αυτο δεν κατεστει δυνατό καθώς δεν έχουν τηλέφωνο αλλά ούτε απαντούσαν στα email που έστελνα . Στις 24/5 μετά απο πολλά μηνύματα με πήρε τηλέφωνο ο κ. Δ και μου είπε πως τελικά δεν πρόλαβα την προσφορά ενώ στην αρχική επικοινωνία που είχαμε που ζητούσε τα λεφτά ήταν όλα οκ. Στο τέλος του είπα αφού δεν είχε το προϊόν να μου επιστρέψει τα λεφτά και εκείνος με τη σειρά του δέχτηκε . Του έστειλα τον αριθμό της τράπεζα για να μου επιστρέψουν τα χρήματα και μου είπε οτι το αργότερο το επόμενο πρωί θα μου επέστρεφαν τα χρήματα . Μέχρι τωρα δεν έχει γίνει καμία επιστροφή και δεν ανταποκρίνεται το κατάστημα σε καμία επικοινωνία .

- 2nd step: Training of the algorithm

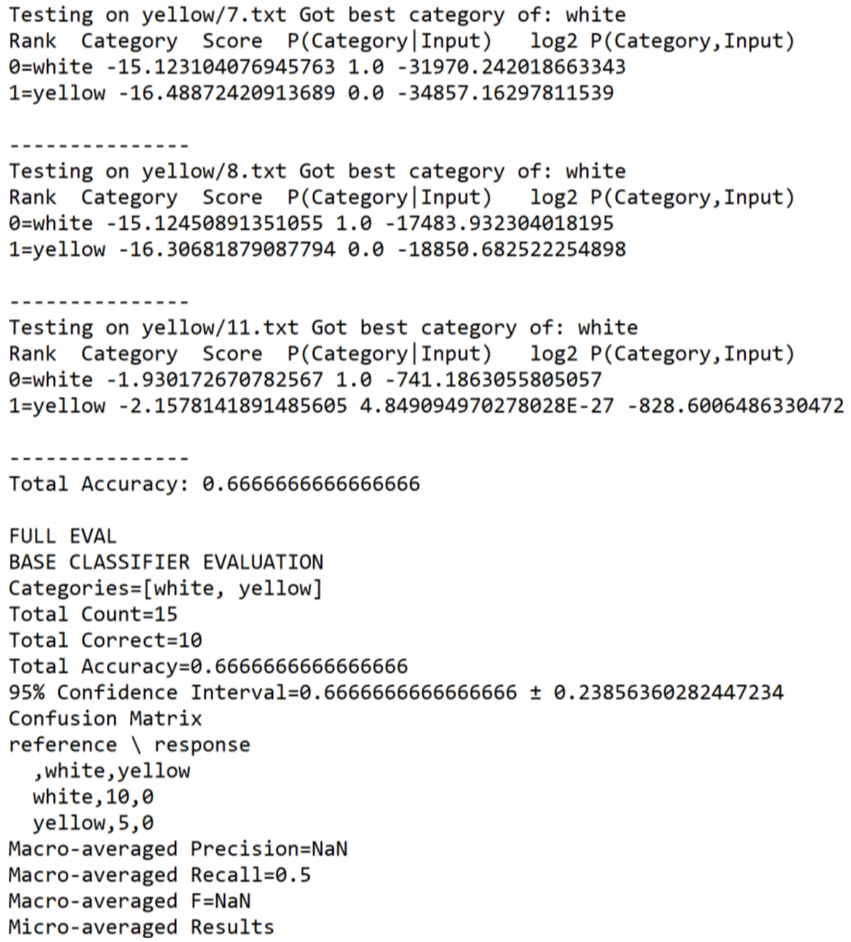
Having completed that step, it is evident that we need to train and optimize the text mining algorithm, in order to minimize the difference between the results. In more detail, we will exploit the qualitative analysis already performed in a specific set of the data (reviews), in order to identify specific problems in terms of logistics operations, and their exact correlation with specific keywords. This process will create the necessary training sets that will be used in order to train the algorithm to identify more efficiently the exact problems and their probability. Using the previous example, we will identify all these reviews that have to do with problematic procedures in terms of refunds when a product is returned. We will train the algorithm to match these and similar reviews with that specific problem, regardless if these two words (return and money) are present in the reviews. Based on that training, the algorithm will then be able to come up with accurate results about the probability of each identified problem, without qualitative analysis needed anymore. In other words, the text mining results will merge with the outcomes of the meaning mining process. For that reason we use the LingPipe Classifier.



```
Training on yellow/51.txt
Training on yellow/35.txt
Training on yellow/14.txt
Training on yellow/39.txt
Training on yellow/22.txt
Training on yellow/25.txt
Training on yellow/28.txt
Training on yellow/59.txt
Training on yellow/20.txt
Training on yellow/54.txt
Training on yellow/44.txt
Training on yellow/42.txt
Training on yellow/71.txt
Training on yellow/21.txt
Training on yellow/72.txt
Training on yellow/38.txt
Training on yellow/61.txt
Training on yellow/32.txt
Training on yellow/80.txt
Compiling
Testing on white/17.txt Got best category of: white
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=white -2.283865193452217 1.0 -5823.856243303153
1=yellow -2.786615129410337 0.0 -7105.86857999636

---------------
Testing on white/2.txt Got best category of: white
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=white -15.140593787765289 1.0 -7434.031549792757
1=yellow -16.575247847839094 8.910712009640681E-213 -8138.4466932889945

---------------
Testing on white/10.txt Got best category of: white
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=white -2.1318074304339665 1.0 -5327.386768654482
1=yellow -2.6914006143096105 0.0 -6725.810135159717
```



```
Testing on yellow/7.txt Got best category of: white
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=white -15.123104076945763 1.0 -31970.242018663343
1=yellow -16.48872420913689 0.0 -34857.16297811539

---------------
Testing on yellow/8.txt Got best category of: white
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=white -15.12450891351055 1.0 -17483.932304018195
1=yellow -16.30681879087794 0.0 -18850.682522254898

---------------
Testing on yellow/11.txt Got best category of: white
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=white -1.930172670782567 1.0 -741.1863055805057
1=yellow -2.1578141891485605 4.849094970278028E-27 -828.6006486330472

---------------
Total Accuracy: 0.6666666666666666

FULL EVAL
BASE CLASSIFIER EVALUATION
Categories=[white, yellow]
Total Count=15
Total Correct=10
Total Accuracy=0.6666666666666666
95% Confidence Interval=0.6666666666666666 ± 0.23856360282447234
Confusion Matrix
reference \ response
  ,white,yellow
  white,10,0
  yellow,5,0
Macro-averaged Precision=NaN
Macro-averaged Recall=0.5
Macro-averaged F=NaN
Micro-averaged Results
```

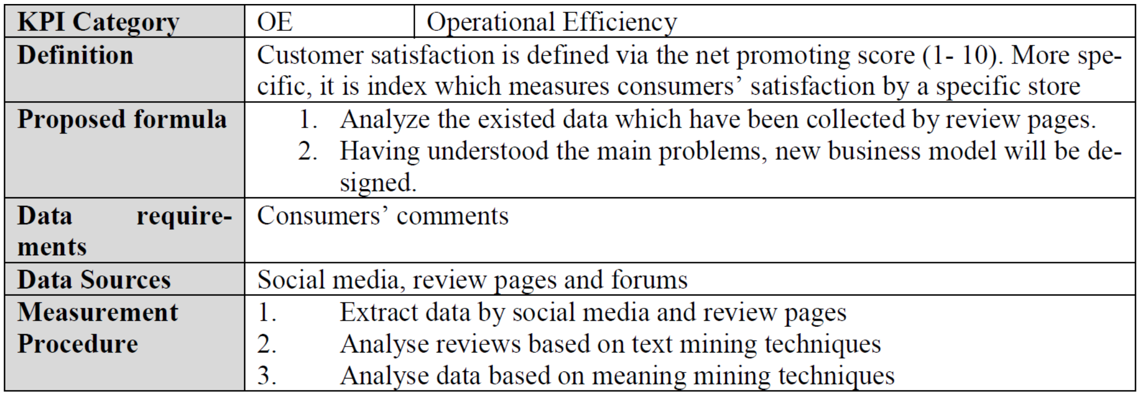
The efficiency of the Classifier was proven to be 66%, meaning that more data may be required to train the algorithm and increase its efficiency.

The Importance of Training Data for Machine Learning

An important part of this process is the data used to train an algorithm. Training data is essentially a set of examples that is given to a machine learning algorithm to "teach" it what to look for. For example, if you want your algorithm to learn to recognize the difference between dogs and cats, you could give it examples of dogs and examples of cats and tell it which was which. The algorithm then uses all of these examples to figure out what features are important in distinguishing between the two classes. The algorithm can then look for similar features in future data to classify it. Thus, training data is extremely important, and if it is incorrect in some way it will have a huge impact on your results. In this example, say instead of giving the algorithm the correct labels of cats and dogs we mistakenly labeled many pictures of cats as dogs. You can imagine that in that case the algorithm would do a very poor job of distinguishing between the two, and you would end up with many cats classified as dogs, an undesirable result. The quality of the training data thus has large impacts on the quality of the results.

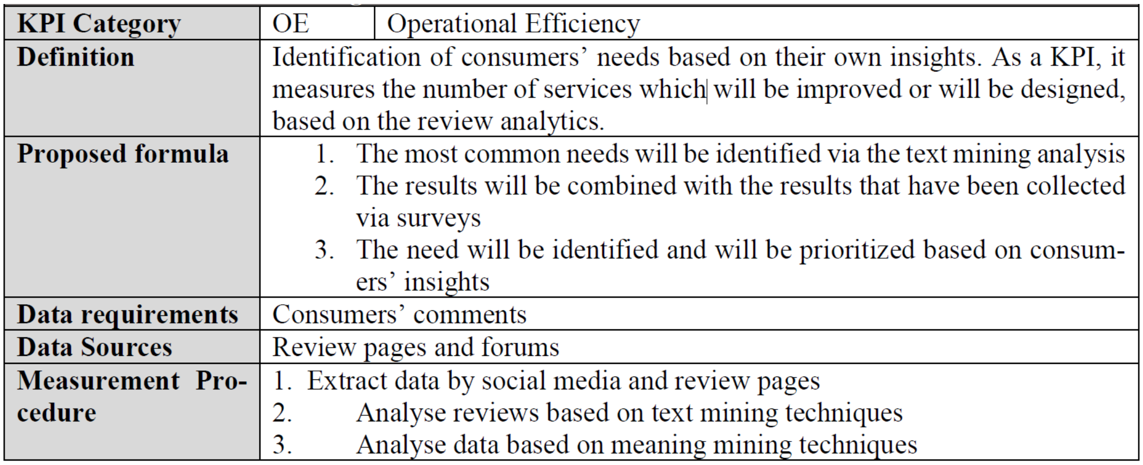
Key Performance Indicator (KPI) Measurement Procedures

Below we describe the how the specific KPIs summarized above related to the improvement of Operational Efficiency (OE) for e-commerce logistics and we provide a structured table on how KPIs will be measured. Improve customer satisfaction with service



KPI Category	OE	Operational Efficiency
Definition	Customer satisfaction is defined via the net promoting score (1- 10). More specific, it is index which measures consumers' satisfaction by a specific store	
Proposed formula	1. Analyze the existed data which have been collected by review pages. 2. Having understood the main problems, new business model will be designed.	
Data requirements	Consumers' comments	
Data Sources	Social media, review pages and forums	
Measurement Procedure	1. Extract data by social media and review pages 2. Analyse reviews based on text mining techniques 3. Analyse data based on meaning mining techniques	

Better understanding of consumers' needs



KPI Category	OE	Operational Efficiency
Definition	Identification of consumers' needs based on their own insights. As a KPI, it measures the number of services which will be improved or will be designed, based on the review analytics.	
Proposed formula	1. The most common needs will be identified via the text mining analysis 2. The results will be combined with the results that have been collected via surveys 3. The need will be identified and will be prioritized based on consumers' insights	
Data requirements	Consumers' comments	
Data Sources	Review pages and forums	
Measurement Procedure	1. Extract data by social media and review pages 2. Analyse reviews based on text mining techniques 3. Analyse data based on meaning mining techniques	

6 Conclusions

The main idea is to embrace key aspects of e-commerce logistics in Smart Cities with various stakeholders' views (customers, online retailers, couriers) and demonstrate how the deployment of Big Data technologies can transform them in order to provide better logistics services and enhance the logistics processes performance.

Following a coherent method of work, we have clarified the current practices and the envisioned ones that rely on collaboration in e-commerce logistics and will, to a large extent, be supported by the big data technologies.

Hence, the aim of this deliverable is to assimilate novel knowledge regarding users' views and requirements for e-commerce logistics in order to frame the implementation and also structure the requirements of the big data infrastructure and the analytics techniques applied.

Our own contribution to this purpose as well as our ultimate goal is to build a system that will identify online consumers' problems and preferences, relevant with logistics procedure, extracting information by social media". Our System will be able to categorize their reviews automatically but mainly find changes in the quota of the problem categories.

This can be used by a logistics companies with thousands of data (posts). Incoming posts will be processed by the Classifier, which will sort them in the corresponding category of problems. So if there is an increase in the proportion of a certain category, this will be an indication that something is wrong, that a problem is growing. Correspondingly, if a decrease is observed in a percentage of a certain problem category, it will be ascertained whether the problem has been addressed successfully.

This tool will be used to attract big industries, in regards to provide reviews from their end users to be analyzed; and hence have another source of big data.

The challenges to accomplish all that are quite a few:

Having the algorithm trained, it will then be able to be utilized in big data, in large amounts of reviews.

1. in huge data, there are several quality issues -> dirty data, missing values, leading to worse training and classification.
2. data velocity - online machine learning requires models to be constantly updated with new, incoming data.

Nevertheless, the key here is to have a tool that will be validated regarding the accuracy of its outcomes and its capability to identify the problems in logistics procedures, as reported by end users. The algorithm will be adopted in order to incorporate the notion of meaning mining and positive reviews will be included in the reviews dataset and further insights will be extracted.

7 References

Chapter 1

1. From the Internet of Computers to the Internet of Things by Friedemann Mattern and Christian Floerkemeier
2. Tutorialspoint - Internet of Things www.tutorialspoint.com

3. Cellary, W. (2013). Smart governance for smart industries. The Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance (ICEGOV '13), October 22–25 2013 (pp. 91–93). Seoul: Republic of Korea.
4. Gil-Garcia, J., Ramon, N. Helbig, & Ojo, A. (2014). Being smart: Emerging technologies and innovation in the public sector. Government Information Quarterly, 31(S1), I1–I8.
5. Meijer, A., & Rodríguez Bolívar, M. P. (2016). Governing the smart city: A review of the literature on smart urban governance. International Review of Administrative Sciences, 82(2), 392–408
6. Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., et al. (2012). Smart cities of the future. European Physical Journal Special Topics, 214, 481–518.
7. Angelidou, M. (2014). Smart city policies: A spatial approach. Cities, 41, S3–S11.
8. International Standards Organization (ISO). (2014a). ISO 37120:2014: Sustainable development of communities—indicators for city services and quality of life. Retrieved August 2016 from https://share.ansi.org/ANSI%20Network%20on%20Smart%20and%20Sustainable%20Cities/ISO%2B37120-2014_preview_final_v2.pdf
9. International Standards Organization (ISO). (2014b). Smart cities Preliminary Report 2014. Retrieved May 2016 from http://www.iso.org/iso/smart_cities_report-jtc1.pdf
10. International Standards Organization (ISO). (2016). Sustainable development in communities. Retrieved August 2016 from http://www.iso.org/iso/iso_37101_sustainable_development_in_communities.pdf
11. Schaffers, H., Komninos, N., Pallot, M., Trousse, B., Nilsson, M., & Oliveira, A. (2011). Smart cities and the future internet: towards cooperation frameworks for open innovation, the future internet. Lecture Notes in Computer Science, 6656, 431–446.
12. Lee, J. H., Hancock, M. G., & Hu, M.-C. (2014). Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco. Technological Forecasting and Social Change, 89, 80–99.
13. International Telecommunications Union (ITU). (2014a). Smart sustainable cities: An analysis of definitions. Retrieved May 2016 from www.itu.int/en/ITUT/focusgroups/ssc/Documents/Approved_Deliverables/TR-Definitions.docx
14. International Telecommunications Union (ITU) (2014b). Setting the framework for an ICT architecture of a smart sustainable city. Retrieved May 2016 from

http://www.itu.int/en/ITUT/focusgroups/ssc/Documents/website/web-fg-ssc-0345-r5-ssc_architecture.docx

15. British Standards Institute (BSI). (2014). PAS 180 Smart City Framework Standard. Retrieved July 2016 from http://www.bsigroup.com/en-GB/smart-cities/Smart-Cities-Standards-and-Publication/PAS-180-smart-cities-terminology/
16. Yigitcanlar, T., & Lee, S. H. (2014). Korean ubiquitous-eco-city: A smart-sustainable urban form or a branding hoax? Technological Forecasting and Social Change, 89, 100–114.
17. Anthopoulos, L., Janssen, M., & Weerakkody, V. (2016). A Unified Smart City Model (USCM) for smart city conceptualization and benchmarking. International Journal of e-Government Research, 12(2), 76–92.
18. Alcatel-Lucent Market and Consumer Insight Team. (2012). Getting smart about smart cities understanding the market opportunity in the cities of tomorrow. Retrieved August 2016 from http://www.tmcnet.com/tmc/whitepapers/documents/whitepapers/2013/7943-alcatel-lucent-gettingsmart-smart-cities-recommendations-smart.pdf
19. Anthopoulos, L., & Fitsilis, P. (2014). Smart cities and their roles in city competition: A classification. International Journal of Electronic Government Research (IJEGR), 10(1), 67–81.
20. Van Bastelaer, B. (1998). Digital cities and transferability of results. In The Proceedings of the 4th EDC conference on digital cities, Salzburg, pp. 61–70.
21. Edvinsson, L. (2006). Aspects on the city as a knowledge tool. Journal of Knowledge Management, 10(5), 6–13.
22. Yigitcanlar, T., O'Connor, K., & Westerman, C. (2008). The making of knowledge cities: Melbourne's knowledge-based urban development experience. Cities, 25, 63–72.
23. Anthopoulos, L., & Fitsilis, P. (2013). Using classification and roadmapping techniques for smart city viability's realization. Electronic Journal of e-Government, 11(1), 326–336.

Chapter 2

24. Thomas Hanne, Rolf Dornberger, 2017, Computational Intelligence in Logistics and Supply Chain Management.pdf
25. Yung-yu Tseng et al. 2004, The role of transportation in logistics chain.pdf
26. Smart logistics van Woensel, T. Published: 01/2012
27. Stephen Ezell, January 2010 - Intelligent transportation systems

Chapter 3

28. Ecommerce Europe. (2016). European B2C E-commerce Report - Facts, Figures, Infographic & Trends of 2015 and the 2016 Forecast of the European B2C E-commerce Market of Goods and Services. Retrieved from https://www.ecommerce-europe.eu/app/uploads/2016/07/European-B2C-E-commerce-Report-2016-Light-Version-FINAL.pdf
29. Glaxon SmithKline Annual Report (2016), Measuring Europe in eCommerce
30. ELTRUN. (2016). Mapping the e-commerce logistics scene.
31. ARCEP (2012), Observatoire annuel des activités postales en France, année 2011. Paris: ARCEP, 36p. Tagesschau (2012). ARD und ZDF Onlinestudie. http://www.ard-zdf-onlinestudie.de/ ATKearny (2012). Von B2C zu B2B durch alternative Zustelloptionen. Aktuelle Herausforderungen für Paketdienstleister im B2C Segment. Page2. http://www.atkearney.de/documents/856314/1214708/BIP_Von_B2C_zu_B2B_durch_alterna tive_Zustelloptionen.pdf/892541a9-63aa-434d-b2b3-2f146094409f
32. Morganti, E., Seidel, S., Blanquart, C., Dablanc, L., & Lenz, B. (2014). The Impact of E-commerce on Final Deliveries: Alternative Parcel Delivery Services in France and Germany. Transportation Research Procedia, 4, 178–190. https://doi.org/10.1016/j.trpro.2014.11.014
33. Esper, T. L., T. D. Jensen, F. L. Turnipseed and S. Burton (2003). "The last mile: an examination of effects of online retail delivery strategies on consumers." Journal of Business Logistics 24(2): 177-203.
34. Agatz, N. A., M. Fleischmann and J. A. Van Nunen (2008). "E- fulfillment and multi-channel distribution–A review." European Journal of Operational Research 187(2): 339-356.
35. *Bask, Lipponen, and Tinnilä, (2012).* E-Commerce Logistics: A Literature Research Review and Topics for Future Research. International Journal of E-Services and Mobile Applications, 4(3), 1-22, July-September 2012.
36. Akter et al., (2016). Big data analytics in E-commerce. A systematic review and agenda for future research. Electron Markets, 26, 173-194, 2016.
37. Koutsabasis, P., Stavrakis, M., Viorres, N., Darzentas, J. S., Spyrou, T., & Darzentas, J. (2008). A descriptive reference framework for the personalisation of e-business applications. Electronic Commerce Research, 8, 173–192.
38. Mehra, G., (2013). 6 uses of big data for online retailers, Practical Ecommerce. Available at: http://www.practicalecommerce.com/ articles/3960-6-Uses-of-Big-Data-for-Online-Retailers (Accessed 2nd of March, 2016).
39. Jeseke, M., Grüner, M., & Weiß, F. (2013). Big data in logistics: A DHL perspective on how to move beyond the hype. White Paper; DHL Customer solutions &

innovation, presented by Martine Wegner, Vice President Solutions & Innovation, 53844 Troisdorf, Germany.

40. Choi, T.-M (2016). Incorporating social media observations and bounded rationality into fashion quick response supply chains in the big data era. Transport. Res. Part E (2016), http://dx.doi.org/10.1016/j.tre.2016.11.006
41. Punakivi, M., Yrjo, H., and Holmstro, J. (2001). Solving the last mile issue: reception box or delivery box? *International Journal of Physical. Distribution & Logistics. Management 31*(**6**), 427–439
42. Pillac,V., Gendreau, M., Guéret, C., &Medaglia, A.L., (2013). A review of dynamic vehicle routing problems. *European Journal of Operational Research 225*(**1**) 1-11. http://dx.doi.org/10.1016/j.ejor.2012.08.015
43. Fabian, J. & Christian, D. (2012). Vehicle routing for attended home delivery in city logistics. *Procedia: Social and Behavioural Science*s (**39**), 622–632. doi:10.1016/j.sbspro.2012.03.135
44. Du, T. C., Li, E. Y., & Chou, D. (2005). Dynamic vehicle routing for online B2C delivery. *The international journal of Management Science* (***33****)*, 33–45. doi:10.1016/j.omega.2004.03.005
45. Gevaers, R., Van de Voorde, E. & Vanelslander, T. (2009). Characteristics of innovations in last mile logistics. *Department of Transport and Regional Economics* - University of Antwerp, 1–21.
46. Duin, JHR; De Goffau, W; Wiegmans, B; Tavasszy, LA; Saes, M (2016). Improving home delivery efficiency by using principles of address intelligence for B2C deliveries. Transportation Research Procedia 12 (2016) 14 – 25
47. Zeiler, K. (2013). Big data in logistics: A DHL perspective on how to move beyond the hype. DHL Customer Solutions & Innovation, 53844 Troisdorf, Germany.
48. Tyan, J. C., Wang, F.-K., & Duc, T. C. (2003). An evaluation of freight consolidation policies in global third party logistics. *Omega*, *31*(1), 55–62. doi:10.1016/S0305-0483(02)00094-4
49. Danloup et al., 2014. Improving sustainability performance through collaborative food distribution International Conference on Green Supply Chain, France.
50. Wang, Gunasekaran, Ngai and Papadopoulos (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications Int. J. Production Economics 176 (2016) 98–110
51. Sage, (2013). Better inventory management: Big Challenges, Big Data, Emerging So- lutions. 〈http://na.sage.com/media/site/erp/responsive/resources/Sage-ERP-Bet ter-Inventory-Management-wp.pdf〉 (accessed 20.04.15).
52. Xu J., Jiang L., Wang S. (2014) Construction of Pick-Up Points in China E-commerce Logistics. In: Zhong S. (eds) Proceedings of the 2012 International Conference on Cybernetics and Informatics. Lecture Notes in Electrical Engineering, vol 163. Springer, New York, NY

53. McKinsey global institute report (2013). Perspectives on retail and consumer goods
54. Gulati and Garino (2000). Get the Right Mix of Bricks and Clicks. Harvard business review, May-Lune 2000 issue.
55. Tanskanen, K., Yrjölä, H., & Holmström, J. (2002). The way to profitable Internet grocery retailing - six lessons learned. *International Journal of Retail & Distribution Management*, *30*(4), 169–178. doi:10.1108/09590550210423645
56. Yu, Wang, Zhong, and Huang (2016). E-commerce Logistics in Supply Chain Management: Practice Perspective Procedia CIRP 52 (2016) 179 – 185
57. Addo-Tenkorang, R, and P. T. Helo (2016). Big data applications in operations/supply-chain management: A literature review. Computers & Industrial Engineering 101 (2016) 528–543.

4th Chapter

58. Charu C. Aggarwal ChengXiang Zhai (2012) Mining Text Data

5th Chapter

59. http://hmcbee.blogspot.gr/2017/05/the-importance-of-training-data-for.html
60. http://nlp.ilsp.gr/ws/
61. http://alias-i.com/lingpipe/demos/tutorial/classify/read-me.html

APPENDIX

Offered tool's functionalities

Near real-time Classification Using Apache Spark with Linear SVM

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.

Using Apache Spark in a clustered environment, one can perform near real time classification on a stream of incoming data, on the basis of a training set of data.

Apache Spark

Apache Spark is a framework for performing general data analytics on distributed computing cluster like Hadoop[1]. It provides in-memory computations for increased speed and data process over map-reduce. Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3. The main advantage of Spark is **speed**: it runs programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

MLlib Machine Learning Library

MLlib is a distributed machine learning framework on top of Spark[2]. It implements many common machine learning and statistical algorithms to simplify large scale machine learning pipelines. MLlib also implements Linear SVM.

SVM is a supervised learning algorithm which is used for classification and regression analysis of data-set through pattern matching[3]. General Pattern analysis algorithms study general types of relations in data-sets such as correlations and classifications. **To apply SVN, the data must be represented in vectors with fixed number of coordinates**:

- Scalar values (temperature, humidity, etc.)

[1] https://spark.apache.org/

[2] https://spark.apache.org/mllib/

[3] https://en.wikipedia.org/wiki/Support_vector_machine

- raster images (need to be converted into n-dimensional vector with each coordinate having some numerical value)

Linear SVM search for a hyperplane that symmetrically separates the data points in the training set between different classes. Here this is known as decision boundary, **that separates the space into two halves: one half for class '0', and the other half for class '1'. This explanation applies only to dataset having two data classes.**

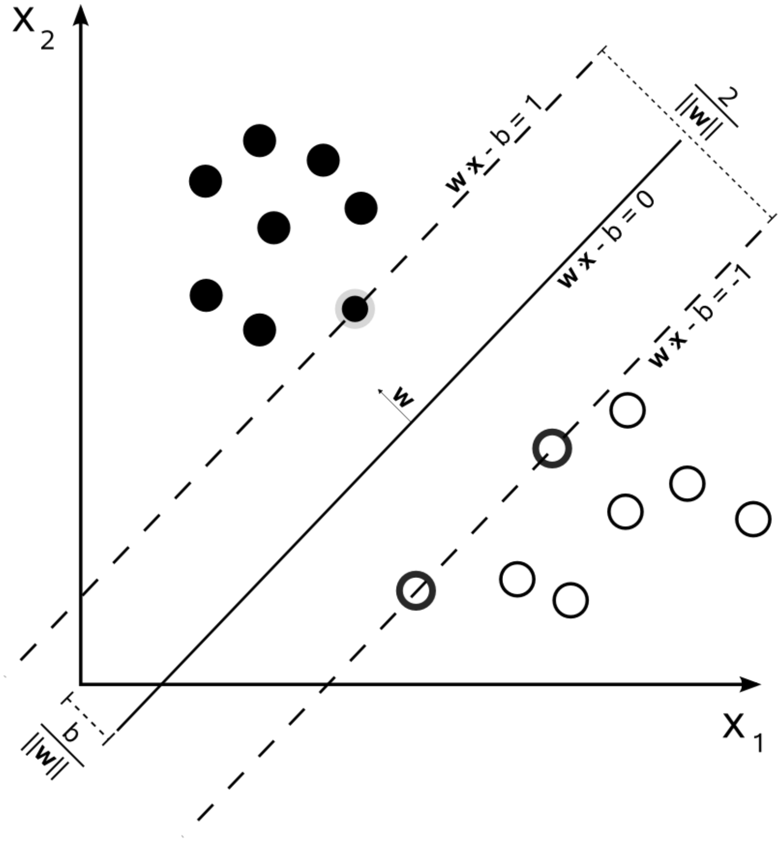

X_2
X_1
$\mathbf{w}\cdot\mathbf{x}-b=1$
$\mathbf{w}\cdot\mathbf{x}-b=0$
$\mathbf{w}\cdot\mathbf{x}-b=-1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

Figure 1 Data points arrangement post linear SVM

Example for Pilot 3.4.1

The pelagic fish stock assessments pilot has described the need for machine learning hybrid analytics: To combine data-driven methods with first principal simulation models and perform artificial dataset synthetization.

The first step is to convert the received data to vectors; that data could be any data related to the specific pilot: images related to fish, sensor data, data related to the vessels' operation etc. Let us assume that the overall target is to classify the data in two categories: Cat1 and Cat2. The vectors can then be used by Apache Spark's MLlib program or Linear SVM as training and test data.

For the training session, one should create two sets (files) of vectors having Class 0 (Cat1) and Class 1 (Cat2).

If classification between more than two classes is needed, we should have many images per category as training dataset. The same process can then be used to classify individual persons, which is practically a face recognition system.