Politecnico di Milano

Faculty of Engineering Department of Electronics, Information and Bioengineering

Master of Science in Computer Engineering

STUDYING THE CORRELATION BETWEEN CONTENTS PUBLISHED ON DIFFERENT SOCIAL NETWORK PLATFORMS

Author:
Deniz Ece Aktan
Matricola:796105

Supervisor:
Prof. Marco Brambilla

Abstract

Studying the correlation between contents published on different Social Network Platforms

In the recent few years, Social Network Platforms have entered into our lives in a revolutionary way. It has been reported in January 2015[1] that approximately more than 2 billion internet users have become active social network users. It has become a new way of communication, which users share the information in many formats, about a variety of subjects. This vast real-time, free formatted data source naturally has taken attention of many individuals and organizations, hence becoming the area of various academic and commercial researches. A great number of works have been conducted to extract useful information from different platforms. On the other side, there are no works focusing on examining the relationship between contents of different platforms. Considering the recent increase in the number of people using multiple platforms, and the lack of work on this subject, the study focuses on examining the relation between multiple social network platforms from the aspect of published contents. In this work different approaches have been proposed and various experiments have been made for giving an explanation concerning this dimension.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

In recent few years Social Network Platforms have entered into our lives in a revolutionary way. It has been reported that by 2015 [1], approximately more than 2 billion internet users have accessed social networks. It has become a new way of communication, which users share information in many formats changing from pure texts to photos, coordinates, links and about many subjects ranging from life events to politics, places, history. As a result of the process, an incredible source of data was produced, where analysis and extraction of information can be made. Naturally this source has taken the attention of advertisers, marketers, politicians, journalists and so on. Creation of this new data source, brought the need to improve the techniques of analyzing the data in these platforms which are completely free in format and limited in length.

One of the most popular social network platforms is Twitter. It is considered to be the leading microblogging social network. According to Alexa reports [12] it is in the top 10 most popular websites since 2013. The IPO filing states that "200,000,000+ monthly active users" access Twitter and "500,000,000+ tweets per day" are posted[13][14]. In this platform users can share maximum of 140 characters that called tweets publicly, or privately to their followers.The users can follow others whose tweets they are interested in. Furthermore they can share pictures, links to websites, and to other social network platform posts.

Another social network platform that is very popular is Instagram. It is used extensively, especially by young population. Instagram has a very fast increase in number of users during recent few years. In December 2014 an announcement made by Instagram blog, stated that [15] it has more than 300 million monthly active users, up from 200 million in March. In this platform users can share pictures, insert a caption, tag other users in the pictures that they share. As in Twitter also in this platform the users can follow the accounts they are interested in, share their posts public and privately according to their preference. In both of these platforms users are able to geotag the post from where it is sent from.

In this part while explaining the concept it is useful to describe some prefixed punctuation symbols that are special to social network platforms. These symbols are used extensively for different reasons, and give a specific meaning about the following text.

To start with, a word prefixed by "#" sign is a hashtag. Hashtags are a way for users to indicate a usually common topic and make the post searchable in public domain about this keyword/ topic. The hash turns any word into a searchable link. This allows people to organize content and track discussion topics based on those keywords. Hashtags are especially important in situations as natural disasters, political events. They are also used while sharing some comments about entertainment, brands and so on to reach many people about a subject. In this work the simplification has been made which assumes that a tweet's hashtag content is a good approximation of its total content [16]. The other punctuation that is fixed for social networks is the @ character. It indicates a social network user mention when @ is followed by a word. There is another format which @ is followed by a space and words, this format is usually an indication of a place name. Although this format is not fixed it is possible to encounter this usage in many posts and it can be used for place extraction.

In the process of data extraction these fixed signs can be used. The fixed characters, and formats add value to the 140 character posts hence enable the analyzers to extract more information.

1.2 Problem Statement

As a result of these revolutionary change of behavior in sharing ideas and the emergence of this huge real-time, free formatted data source, social networks have become the area of many academic and commercial researches. A great number of researches have been conducted to extract useful information from social network posts. Concept extraction, entity extraction, sentiment analysis, hashtag suggestion can be given as example studies on these data. In the process of conducting these researches usually NLP techniques and algorithms have been used. These techniques are appeared to be ineffective at analyzing social network posts, which in nature often contain slang, acronyms, or incorrect spelling or grammar.

In contradiction to the high amount of researches that have been made for extracting information from a particular social network, there aren't many studies that conducted between multiple, different formatted, social networks to examine the correlation of user behavior and posts between different platforms.

The report of pew [17] in January 2015 states that according to the results of survey on American adults using the internet "multi-platform use is on the rise: [18] 52% of online adults now use two or more social media sites, a significant increase from 2013, when it stood at 42% of internet users". It can be told that there is an obvious increase in the number of the social network platforms and number of users sharing in multiple platforms. Hence, as well as examining each platform's dynamics, it can be also interesting trying to extract information from different platforms. In the thesis work it is suggested that there may be some correlation of the user posts between different network platforms in terms of time, entity, sentiment, number . This may give us an idea about the motivation of usage, the tendency of sentiments of the posts in some interval of the time or platform, as well as well as the prediction of popular subjects. Interesting results may provide the platform developers useful insights for improving their implementations, give ideas about new social platforms, more efficient user interfaces.

1.3 Proposed Solution

In order to examine the relation between different social network platforms one is required to:

- Decide which platforms to be considered for comparison.
  There are many social networks that can be considered; as Facebook, Linkedin, Twitter, Instagram, Foursquare, Swarm, Flickr, Pinterest. In this study the availability of public posts was the main reason for preference. Facebook, although being the most popular platform, was not selected, because of its limitation of publicly available posts. The same fact holds for Swarm and Foursquare. For this reason Twitter and Instagram, which can provide a great number of publicly available posts, were preferred to be used for following analysis.

- Collect the data.
  The Twitter and Instagram data, which has been collected during previous work [11] [9]is used.The time interval that has been covered in both datasets is identified, the posts that were published in this intersection of time are extracted and cleaned for the use of analysis.

- Evaluate the metadata.
  Data properties are evaluated to identify the possible aspects that the relation can be discovered.
  - Instagram posts contain photos, captions of the photos, geo-tagging, comments, and likes for these photos,
  - Twitter posts contain texts, information about retweets, favorites, geotagging.

  For the ease and the limitations of image tagging libraries and web services, it has been decided to focus on the relations between texts in posts as well as other metadata that the they provide us i.e. hashtags. The entities a Tweet and an Instagram post contain are hashtags, links, user mentions.

- Identify the aspects.
  After the decision to concentrate the work on the textual content has been made, it was crucial to specify the subjects of comparison. It was proposed to
  - Explore topics mentioned and check if there is some relation between the subjects of the posts.
  - Examine the tendency of sentiments in two platforms, check for the relation of sentiments about a specific subject between these platforms.
  - Analyze the mentioned entities i.e. in terms of sentiments, popularity and time and check for the relation.
  - Take different time granularities, and specific intervals into consideration in all experiments.

1.4 Structure of the thesis

The remainder of the document is organized as follows:

- Chapter 2 gives general overview of the relevant concepts and technologies for the thesis project.
- Chapter 3 introduces other works that address similar problems through discussing the associated publications.
- Chapter 4 describes the novel perspective to the analyzing relation of contents published on different platforms and the reasoning behind.
- Chapter 5 describes in detail how the project is implemented.

- Chapter 6 gives the experiments and results acquired from real data, using this implementation.
- Chapter 7 summarizes the work done, the important results obtained, and discusses how the project can be improved.

Chapter 2

Background

2.1 Relevant Concepts

2.1.1 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. [19] KDD is the organized process of identifying valid, novel, useful and understandable patterns from large and complex data sets. In recent years, with the appearance of new social networks and technological advancements collecting data has become easier and storing it has become inexpensive. This created the availability of huge stored data, in different geographical places, amounts and formats. It has been estimated that the amount of stored information doubles every twenty months[20]. The data availability is increasing exponentially while the human processing level remains constant. Hence the gap increases exponentially. This gap is the reason of the need and creation for the KDD field, which, therefore, becomes increasingly important and necessary. Knowledge discovery in databases is a process consisted of a variety of other subprocesses rather than being a simple concept. KDD consists of [21]:

2.1.1.1 Data Cleaning

Incorrect or inconsistent data can affect the analysis process leading to false conclusions, decreasing the quality and precision of the results. Hence one of the first and most important steps in data processing is data cleaning. Before the analysis can start, it is essential to verify that the data values are correct or, at the very least, conform to some a set of rules. Data cleaning aims to remove noise or irrelevant data as well as detecting and cleaning the extreme outliers. Data cleaning itself consists of few subprocesses:

•Data auditing: To detect which kinds of inaccuracies and inconsistencies are to be removed, a detailed data analysis is required. As well as statistical and database methods also the manual inspection may need to be used to identify data quality problems and metadata about data properties.

•Definition for Workflow specification and mapping rules: Depending of the number of the data-sources, data structures and the variety of the inconsistencies, inaccuracies there may be many steps to be executed and mapping rules to be identified on the data for sake of increasing the quality. In this step for achieving a proper workflow, the reasons of the anomalies and errors in the data have to be closely considered.

•Workflow execution: In this part the workflow for cleaning the data that has been defined in the previous step has been implemented and executed.

•Controlling: The part that the data that has cleaned is controlled and verified.

2.1.1.2 Data Integration

Data integration is the step where the data from multiple resources, in different formats and properties are processed with the aim of giving the analyzer a unified view, and merged together. This is the step where the metadata for different datasets should be carefully examined and the entity matching and schema integration is conducted.

2.1.1.3 Data Selection

As it can be inferred from the name, in data selection phase the relevant data for the analysis to be followed is identified and this subset of the datasets has been retrieved.

2.1.1.4 Data Transformation

In data transformation step the selected data are transformed or consolidated into forms that are appropriate for mining.[22] Many procedures can be used depending on the data format and analysis in this step;

•Smoothing

•Aggregation

•Generalization

•Normalization

•Attribute Construction

2.1.1.5 Data Mining

Data mining is the extraction of implicit, previously unknown and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data.[23] Data mining uses procedures and techniques from machine learning and artificial intelligence, pattern recognition, statistics and database systems. Data mining can be classified into two categories; descriptive data mining which represents the interesting properties of the data and predictive data mining, which constructs models for making an inference from the existing data for predicting behavior of new datasets.

Here are some major tasks of data mining:

•Class Identification (summarization) : Gives a precise idea about the collection of the data that has been analyzed. Covers the summary properties; count, sum, average as well as the numbers about data dispersion such as variance, covariance, quartiles and so on.

•Classification : Analyzes each set of training data and constructs a model based on the features of these different sets of data. A set of rules or a decision tree is constructed by the process aimed to be used for classifying the future data.

•Association : Examines the association relations and correlations between different sets or subsets of datasets.

•Deviation Detection : Detects significant changes in data from normal behavior.

•Clustering : Identifies the clusters contained in the data, the main principle is to group in a way that maximizes the intraclass similarity and minimizing the inter-class similarity.

•Time series analysis: Analyzing large set of time-series data to discover some patterns, interesting regularities, similar sequences.

•Association Rule Discovery: Finds interesting association/ correlations in huge data sets, derives dependency rules in order to be used for prediction of future values

•Sequential Pattern Discovery: Reveals interesting relationship features depending on the time interval, derives sequential behavior rules for prediction of future data.

2.1.1.6 Pattern Evaluation

This step is the interpretation of the detected pattern to identify if it is unfolding the a truly interesting pattern representing knowledge, based on predefined measures.If the identified pattern is not interesting, the cause for it should be figured out. It will probably be necessary to fall back on a previous step for another attempt.

2.1.1.7 Knowledge Representation

As can be understood from the name, knowledge representation is the step where the visualization and knowledge representation techniques are used to present discovered knowledge and interesting information to users. This step might include providing a graphical user interface that allows the users to browse datasets and data warehouse schemas or data structures, evaluate mined patterns and visualize the patterns in different formats.

2.1.2 Natural Language Processing and Information Extraction

Natural Language Processing is a field of computer science, artificial intelligence and computational linguistics that concentrates on natural language understanding, which involves enabling computers to derive meaning from human text. It has developed various techniques that are typically linguistically inspired, i.e. text is typically syntactically parsed using information from a formal grammar and a lexicon, the resulting information is then interpreted semantically and used to extract information about what was said.[24]

Here are some of the major Tasks of NLP:

•Machine Translation : Concentrates on automatic translation of a text from one language to another.

•Morphological Segmentation : Separates the words and identifies morpheme class for each word depending on the language. Complexity may change according to language.

•Automatic summarization : Provides a summary of a given text based on the extracted key sentences and phrases giving the user a rough idea about the content of the text.

•Named Entity Extraction : Takes a text a part of text as an input and identifies which words or phrases corresponds to an entity. Entity here might refer to a place, person, brand and so on.

•Topic Segmentation and Recognition : Concentrates on the extraction of single or multiple topics that identify a given text input.

•Speech recognition : Takes a sound clip of people speaking and outputs the textual representation of the speech.

•Natural language generation : Concentrates on transforming the information from the computer to the human language.

•Natural language understanding : Enables computers understand the human language converting the natural language to first order logic that a machine can understand.

•Part-of-speech tagging : Concentrates on identifying the part of speech for each word in a given sentence. (Part of speech in this context refers to noun, verb, adjective and so on.)

•Question Answering : Given the question in human language as an input tries to output an answer to this question.

•Coreference Resolution: Given a part of text as an input, concentrates on determining that are referring to same objects.

2.1.3 Crowd Sourcing

The recent revolutionary change of behavior in sharing ideas and massive peak in the social network usage, created a new free potential source for researchers and data analyzers which is derived from the main idea of “wisdom of crowd”. This phrase means that the evaluation of answers from multiple individuals would give a more accurate idea about a question than the answer of an expert.

Deriving from this main idea the questioning has appeared about how to use this new source of the crowd to complete specific tasks. The concept of crowd-sourcing has borned from this idea. The concept of crowd-sourcing is basically, using the community of people that is available in social networks, to complete a job, just like distributed computing concept but in this case with humans instead of computers. An information or input for a particular task is being outsourced by a number of people either paid or unpaid usually via internet.

The history of crowdsourcing comes from other issues: crowdsourcing as a novel methodology for user-centered research; development of new services and applications based on human sensing, computation and problem solving; engineering of improved crowdsourcing platforms including quality control mechanisms. [25] The crowdsourcing is categorized into two types:

•Explicit crowdsourcing: In this type the main aim is the same with the users and they can cooperate to achieve it.

•Implicit crowdsourcing : In this type the main aim is not the same with the users.

Two subcategories can be identified for this type :

•Standalone: the results are a side effect of the task they are actually conducted by the users

•Piggyback: the results consist of the users' data from a third-party website.

2.1.4 Crowd Searching

CrowdSearching is a novel search paradigm that embodies crowds as first-class sources for the information seeking process.

The concept was born to automate the search process that a human usually does, when only mere factual information are not enough. Sometimes a user is more interested in the human insight or opinion, but to do this he's forced to interact physically with friends, or communicate or via messages manually posted on social networking platforms. Then often make up his mind by combining results from search engines, investigations on vertical portals, and opinions gathered within their friends and trusted people circles. [26] If we look better, what we usually do is crowd-sourcing, ask something to the crowd and then combine manually the different opinions. Crowd-search automates the process, integrates results of a conventional search with the information from crowd-sourcing.

2.1.5 Social Media Monitoring

Nowadays the spread of social media and the opinions, comments, feelings that people share using these platforms, create a massive, precious data for companies. They are forced to adopt special methods in order to extract useful data, for being competitive in the market. It can be stated that SMM is the continuous systematic observation and analysis of social media networks and social communities in order to better understand the opinions of users about a product, or a company.

The information that the data encloses is incredibly important by many aspects. From the company point of view, it is possible to identify brand's overall visibility, what the consumer think, what they like or dislike about the company, or about a specific product. They can also calculate some statistics divided for customers segments, by age, gender or location, without the need to ask them, because the major part of social media users have specified all their personal data. These are some opportunities that SMM provides to the companies, all of these belong to the phase of listening, but there are also other opportunity with the second phase of presence, if the company decide to join the social media with creation of its official page or similar, for example interact with the consumers or involve them with challenges (co-creation of a company mark) in order to get the brand stronger and improve clients loyalty. On the other side, there are many problems which make the use of these systems not trivial.

The first problem is the size of the data, there are various social network platforms, with millions of active users that every day publish information on the web. The second problem is the sentiment analysis, the automatic process that starting from a textual comment is able to understand if this comment is positive or negative in a specific context.

The third problem is the spam. Social media is full of data that may be non-relevant, so it's necessary to filter carefully for the sake of precise and accurate results. For all these reasons was born the social CRM, a module that integrate these functionalities for support social media, with the rest of enterprise information system.

2.2 Relevant Technologies

2.2.1 Language Detection : LangDetect[2]

To make the analysis on the text in many aspects, the first thing required is to identify the language of the text. The detection of language is achieved using the library LangDetect. This library is implemented by generating language profiles from Wikipedia abstract xml, using these profiles for detecting language of a text using naive Bayesian filter.

•It is claimed to have 99% over precision for 53 languages The texts of the social network posts are given to the detection function of the library as an input parameter, which returns a result in the form of an array. Each element of the array includes the language code with the probability of the text belonging to this language category.

2.2.2 Sentiment Analysis, Categorization, Entity Extraction Web Service : Aylien[3]

To make the analysis in the aspect of sentiments Aylien web service is used. The texts of the social network posts having the English language have sent to be labeled by the service. The output includes the label for polarity of the sentiment (one of the three from: negative, neutral, positive), the confidence of this label in a number from 0 to 1, the label for subjectivity and the confidence of this label in a number from 0 to 1.

As well as sentiment analysis, for the analysis of the sentiments of the posts mentioning specific entities, the entity extraction functionality of this web service is used. The texts to be extracted or a URL can be sent as a parameter to the service and it returns a structure containing many possible lists; organizations, locations, persons, notable keywords, date values, money values, percentage values, time values, URLs, email addresses, phone numbers that are found in the text are returned. In case of no values found, corresponding to one category that list doesn't exist in the response.

2.2.3 Spanning the time interval : Joda Time[4]

For spanning the time interval in different granularities as month, week, day the existing functionality that is provided by the java Date remained incapable, Joda Time offering a variety of functionalities ranging from Date comparisons in terms of chronology to timezone support, support for different calendars. Joda-Time has become the de facto standard date and time library for Java. From Java SE 8 onwards, users are asked to migrate to java.time (JSR-310). The design allows for multiple calendar systems, while still providing a simple API. The 'default' calendar is the ISO8601 standard which is used by many other standards. The Gregorian, Julian, Buddhist, Coptic, Ethiopic and Islamic calendar systems are also included. Supporting classes include time zone, duration, format and parsing.

2.2.4 Using Mongodb data : JONGO[5]

Mongo driver for java. Provides an API containing functionalities ranging from querying and updating mongo collections to mapping the mongo objects to java objects. Jongo gets access to Mongo with its Java driver and relies upon the Jackson marshaling library well known for its performance to offer the comfortable Mongo shell experience in Java.

2.2.5 Evaluating the correlation : Apache commons' Math Library[6]

For evaluation of the correlation between Instagram and Twitter; the results which are acquired as Hashmaps needed to be processed and evaluated. Apache commons math library's statistics package is used to calculate the correlation coefficients for both Spearman and Pearson values.

This package includes a framework and default implementations for the following Descriptive statistics:

•Arithmetic and geometric means

•Variance and standard deviation

•Sum, product, log sum, sum of squared values

•Minimum, maximum, median and percentiles

•Skewness and kurtosis

•First, second, third and fourth moments

With the exception of percentiles and the median, all of these statistics can be computed without maintaining the full list of input data values in memory. The stat package provides interfaces and implementations that do not require value storage as well as implementations that operate on arrays of stored values.

Chapter 3

Related Work

This chapter introduces other works that address similar problems through discussing the associated publications.
The overall popularity of social network platforms has created various research themes. A great number of studies with regard to the content, the dynamics and the structural characteristics of Twitter, Facebook, Foursquare, has been conducted in recent years. There are works which concentrate on event extraction[8][10], individual tracking[10], topic identification and clustering of hashtags[7], the information extraction about mobility[11] as well as dynamics of interest between cities[9]. These works have brought new perspectives and solutions to Social Network Analysis area and created the basis for the future works on information extraction from social media platforms. On the other side, the researches are concentrated on one specific social network, especially on Twitter, and the relations between multiple platforms didn't catch any attention. The work discussed in this paper, as well as grounding the basis on the related work explained in this chapter, brings a new perspective for studies concentrating on social network analysis.

3.1 Exploring the Meaning behind Twitter Hashtags through Clustering [7]

This paper concentrates on retrieving connections that might exist between different hashtags and their textual representation, as well as grasping their semantics through the main topics they occur with.

The proposed solution was based on the idea that the hashtags in tweets are a good representation of their contents. Basing on this idea, the work focusses on analyzing and representing hashtags by the corpus in which they appear. A large set of hashtags is clustered using K-means on map-reduce in order to process data in a distributed manner.

It is assumed that each hashtag has a unique representation in the dataset, composed of the concatenation of all tweets which include it. It is hypothesized that encountering two different hashtags to have the same so-called virtual document is low in probability. In other words, it is improbable that considering a couple of specific tags, every time one of them occur in one document the other tag will be mentioned in the same document. 280.000 distinct hashtags from 900.000 daily tweets per dataset are clustered using K-means while varying the number of clusters k. K represents the category quantity.

Approach followed, consisted of multiple steps:

- Data Cleaning:
  - The tweets that are not containing hashtags are removed.
  - Tweets in English are kept.
- A top 10 of most frequent hashtags per dataset and their corresponding frequency is examined:
  - Almost none of the hashtags follows the classical pattern of tagging with terms.
  - Some of them include abbreviations and terms of concatenated words.
- The hashtags are represented by their documents
  - It is shown that they follow a power law, while a few popular hashtags repeat themselves in the collection a great number of times, a large number of hashtags have a small frequency. Translated in the dataset, popular hashtags are represented as rich documents, while very less frequent ones have poor documents. This results with a sparse vector.

It can be seen that most tweets have a small number of hashtags, and just a few, a large number. After manual observations, it has been concluded that these type of tweets represent spam. Users put together several popular hashtags and a shortened URL in order to drive traffic to a web page.

- Texts are cleaned from mentions and URLs,
- Virtual documents are created for each hashtag, consisting of the concatenation of all the tweets in which it was mentioned.

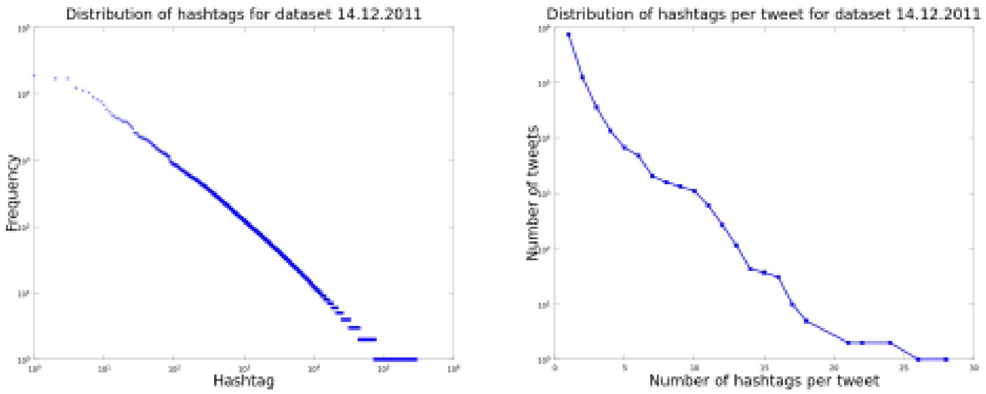
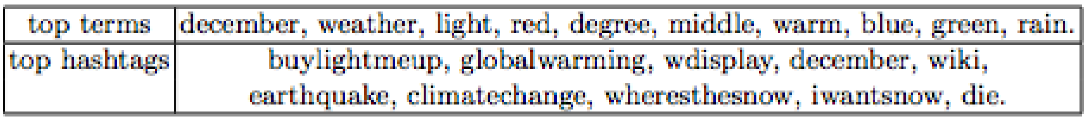


FIGURE 3.1: Results of proposed Approach

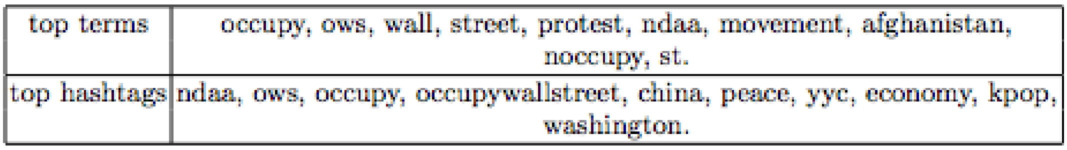
- Stop words are identified and eliminated, rare words and spelling mistakes are removed with the selection of a minimum frequency of 10.
- K-means clustering algorithm has been applied on the dataset for categorizing the tweets. The results of the clustering show that it is possible to identify semantically related hashtags. For each cluster the top terms were extracted, i.e. the most frequent terms in the virtual documents of the cluster. These top terms are the most representative for the cluster, and fulfill their role as explanatory terms.

  Furthermore, top hashtags within a cluster were extracted, and they are obtained by ranking all the hashtags in the cluster by an importance score. This score is computed multiplying the centrality of the hashtag, i.e. the distance from the centroid, by the dimension of its virtual document, that is proportional to the popularity of that hashtag.



top terms	december, weather, light, red, degree, middle, warm, blue, green, rain.
top hashtags	buylightmeup, globalwarming, wdisplay, december, wiki, earthquake, climatechange, wherestheснow, iwantsnow, die.

FIGURE 3.2: Results with k=100



top terms	occupy, ows, wall, street, protest, ndaa, movement, afghanistan, noccupy, st.
top hashtags	ndaa, ows, occupy, occupywallstreet, china, peace, yyc, economy, kpop, washington.

FIGURE 3.3: Results with k=500

The idea of simplifying tweet contents' to hashtags, the preprocessing steps E.g. data cleaning, extraction, and the main notion of the approach was used as a guide during the thesis study.

3.2 Open Domain Event Extraction from Twitter [8]

In this paper the "twitCal", the first open domain event extraction and categorization system is explained. Accurate extraction of the open domain calendar of significant events is feasible to achieve from Twitter, by this approach.

Previous work in event extraction has focused largely on news articles, as historically this genre of text has been the best source of information on current events. The disorganized structure of tweets is motivating the need for automatic extraction, aggregation and categorization.

TwitCal extracts a 4-tuple representation of events which includes a named entity, event phrase, calendar date, and event type. The steps of the approach are explained below:

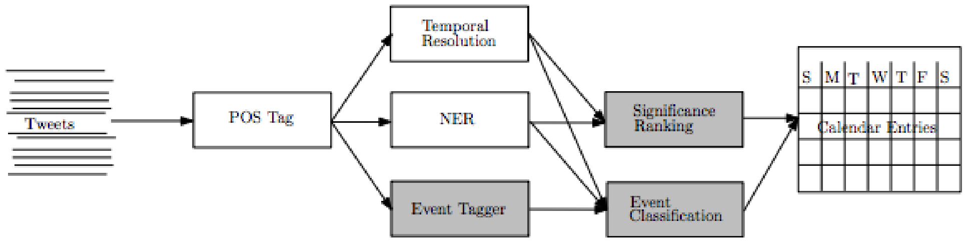


FIGURE 3.4: Steps of Approach

- First the tweets are POS tagged : A pos tagger is used that is tuned for Twitter posts is used.
- Named entities and event phrases are extracted : In this step an entity extraction trained on the Tweet posts is used [27]

  Event phrases can consist of many different parts of speech as below;

  - Verbs: "Apple to Announce iPhone 5 on October 4th"
  - Nouns: "iPhone 5 announcement coming Oct 4th"
  - Adjectives: "NEW IPHONE TODAY!"
- Temporal expressions resolved: In general there are many different ways users can refer to the same calendar date, for example 'next Friday', 'August 12th','tomorrow' or 'yesterday' could all refer to the same day, depending on when the tweet was written. To resolve temporal expressions TempEx was used.
- Extracted events are categorized into types: An approach based on latent variable models infers an appropriate set of event types to match the data. Furthermore, these events are classified into types by leveraging large amounts of unlabeled data.

- Evaluation of association : Strength of association between each named entity and date based on the number of tweets they co-occur in, was examined, in order to determine whether an event is significant.
- The ranking has been done according to significance values.

Top 100,500, 1000 entries are evaluated concerning different aspects giving the precision higher than 0.7 for top 100 entries.

The main idea of this paper; extraction of entities from social network posts, and the approach which utilizes time for identifying the events was used as a guide in the thesis study.

3.3 Relevance, Relations and Topics of Interest in Cities Based on Content Mining from Social Networks [9]

This paper is a thesis study from Politecnico di Milano, focused on classifying and geotagging the social network posts and examine the relations between cities based on the social media posts.

Project can be examined in 3 phases,

1) Extraction : The posts limited to relevant cities have determined and extracted from Twitter and Facebook. By geolocation based search and keyword based search i.e. searching by city names that are interested in.

2) Transformation:

•Language detection step : For Twitter posts this is done by using the API, for Facebook posts this is done by using Language detection library (Google). After the detection of language only English and Italian posts are kept, the others are translated to English (Yandex web service)

•Categorization step: In this step, 2 classifiers are used; one with the training test composed only by posts in English and another one with training test composed only by posts in Italian.

•Hashtags step : In this step hashtags are taken as an array from tweets

•Place detection step : Geolocation or dandelion & mapquest

3) Network analysis step : Source and destinations are identified.(Source:place of the author, Destination: place of the post) Load: Store the processed data,

After the data is processed and stored as needed it is analyzed and demonstrated visually through a web application

Different types of visualization are:

•Mentions map: a map that shows links, in terms of mentions between different places, with arrows. The starting point is the source place, that mention the place pointed by the arrow.

•Category pie chart: A pie chart that shows the different weights that the different categories have. It's possible to filter by city, or see it on all the dataset.

•Mentions column charts: Two column chart, the first shows the main five countries that are the source of mentions towards one city.

•Hashtags list: a list that shows the thirty most popular hashtags, for all the dataset, or by filtering on a specific city.

My thesis study draws on many ideas from this work, as steps of approach, the concept of examining the contents mentioning some specific cities. Furthermore, the result dataset of this work was used as a source.

3.4 Timeline Generation: Tracking individuals on Twitter [10]

In this paper the research and experiments based on constructing a user's important life events from his/her Twitter timeline are discussed.

An unsupervised framework is proposed for extracting a user's important life events.

The events are classified according to two features

•Time specific vs Time General

•Public vs Personal

So ended up with 4 classes:

•Public Time General Events

•Public Time Specific Events

•Personal Time General Events

•Personal Time Specific Events

It is analyzed that for ordinary Twitter users personal time specific events are the ones that can be classified as Personal Important Events (PIE). On the other side for celebrities the situation is different since for a basketball player for example, a tweet about a game can also be categorized as a personal important event. So for increasing the accuracy different approaches for extractions of PIE's for celebrity and Ordinary people is used.

Three criteria for PIE extraction is introduced.

- First, a PIE should be an important event, an event that is referred to many times by an individual and his or her followers.
- Second, each PIE should be a time-specific event, a unique (rather than a general, recurring and regularly tweeted about over a long period of time) event that is covered especially by specific start and end points
- Third, the PIEs identified for an individual should be personal events (i.e. an event of interest to himself or to his followers) rather than events of interest to the general public.

  Process consists of multiple steps :
- Topic merging : Hierarchical clustering algorithm is employed on the topics that were extracted by means of Dirichlet Process Mixture Model. Closest topics were merged into a new one, step by step until the stopping conditions are met.
- Celebrity related public Time Specific identification : To identify celebrity related Public Time Specific topics, rules are employed based on
  1. Username co-appearance
  2. p-value for topic shape comparison
  3. Clustering balance are conducted.
- Gold-Standard Dataset Creation
  1. For ordinary users PIE labeling : 20 users was asked to identify each of his or her tweets as either PIE related according to their own experience. (Twitset-O)
  2. For celebrity users labeling : Amazon's Mechanical Turk used to label celebrity tweets. (Twitset-C)

In the experiments for modeling

- Dirichlet Process Mixture Model
- Multilevel LDA
- Person DP: A simple version of DPM model that takes only as input tweet stream published by one specific user
- Public DP: A simple version of DPM model that takes only as input tweet stream published by one specific user. has been used the results were pretty satisfying with the recall rate of 0.9 with DPM for Twit-C and rate of 0.7 with Public DP for Twit-O.

The approach and experiments explained in this paper and the notion of categorizing the extracted entities considering time, was utilized in the main idea of the thesis work.

3.5 Monitoring Urban Mobility With Social Networks[11]

A previous thesis work focusing on the real-time traffic monitoring using the data extracted from social network platforms. There are existing services, that offer information about mobility, on the other side they are missing personal data which may provide more detailed insight. In addition the information they use is taken from the community of their users, this fact causes the information not to be reliable or sufficient for providing the service. For this reason in the work the idea is to use an existing community as Twitter and Facebook which are extensively used by the people to extract the information about the mobility as well as providing personal insight. The project tested in Milano city. Consisted of 3 main stages

1) Extraction: Posts from Twitter and Facebook are extracted using Twitter4j and Restfb

2) Classification: Lingpipe web service is used.

As the initial step, the mobility related and non-mobility related posts are identified, unrelated ones are eliminated.

To do this, the system classifies all the posts regarding the mean of transport to which they refer using LingPipe. The monitoring for a wide range of transport has been supported by this work; train, car, airplane, bus, subway, tram.

After the classification step the detailed categories are identified (On car and train):

Train related posts are further classified as:

•Train with fault;

•Train dropped;

•Train delayed;

•Train crowded.

Car related posts are further classified as:

•Traffic;

•Accident;

•Road closed;

•Difficult driving conditions;

•Work in progress.

3) Geotagging : As many of the posts lacks geolocation, the non-existing geolocations are acquired by means of a named entity extraction software 'Dandelion'. After identifying the name of the place, the coordinates of the place are obtained using MapQuest web service.

In the following step, this information is provided as a web service and integrated with Infoblu's official web service. The results are shown with a web app, that provides to the users multiple visualization opportunities:

•All the messages in a list;

•All the messages positioned on a map

The main idea of this work and the steps of approach was used as a guide in my thesis. In addition the result dataset of this work was used as a source.

Chapter 4

Analysis of Correlation between contents published on different Social Network Platforms

In this section of the document, the main idea behind the provided solution will be discussed in more detail. The aim is to go deep into the analysis approach, explaining in detail each step, what is needed and why.

4.1 Main Idea of the Study

The experiments and work that have been done, grounded on the idea that there might be a relation between contents published on different platforms, in terms of time, popularity and sentiments. To examine the strength of the relation in these aspects, the entities contained in the tweets and the textual contents have taken into consideration. The hashtags, geotaggings, places mentioned, and the creation dates were proposed to be used for the research. Especially the timestamps of the posts were used in each analysis step, to get a result that is time-related. Furthermore to get an extensive idea about the relation, it was proposed to analyze the tendencies of the posts by means of sentiments in general, as well as sentiments of the posts mentioning specific subjects. Taking these proposals and ideas into consideration, the details of the research was identified, in terms of structure needed to be examined and aspects to be considered.

4.1.1 Topic Correlation

To identify if there is some correlation between different platforms the first aspect to be examined was the topics. For examining the posts in terms of topics, there were multiple options:

- To classify the texts of the posts, using a web service or NLP tool. Check after, if the category labels of the posts were correlated.
- To use entity extraction on the social network posts and identify if there is a relation between these 2 different social networks in the matter of the entities that are referred in the posts.
- To examine the hashtags and make the simplifying assumption that the hashtags give an idea about the social network posts' topic.

At this step as it has been observed that the NLP tools don't provide accurate results for categorization of social network posts, because of their limit of length, and the language which is completely free in structure (including slangs, without grammar rules). For this reason, the added values as fixed punctuations were decided to be used, for labeling the topic of the tweets. The hashtags, the words that are coming after the hash sign, are used usually for manually labeling the post as a specific common topic by the user itself. The simplification has been made, assuming that the hashtags' content is a good approximation of the posts total content [16].

Since hashtags are vastly used in the posts in both of the datasets, Twitter and Instagram, this simplification helped a lot to omit the categorization part on post texts which may cause erroneous results hence affecting the evaluation of correlation . The API which is provided by the Twitter and Instagram in the structure of the json returned from the web service include the hashtags as an array.

4.1.2 Sentiment Correlation

The next aspect to be considered for examination of the relationship was the sentiments. The process of searching for a relation between the sentiments of the posts through social network platforms consisted of multiple analysis explained below.

- The tendency of sentiments in general were analyzed, between Twitter and Instagram by means of polarity and subjectivity, searching for a specific behavioral pattern.

- The next analysis was focused on the sentiments that are extracted about the entries referencing the top 100 topics in a time interval. The posts belonging to the topic of interest were examined to get the sentiment. Afterwards the analysis was made to check if there is a consistency in the sentiments for each topic between two different social network platforms in weekly time intervals through the datasets.

4.1.3 Sentiment analysis for 4 big cities

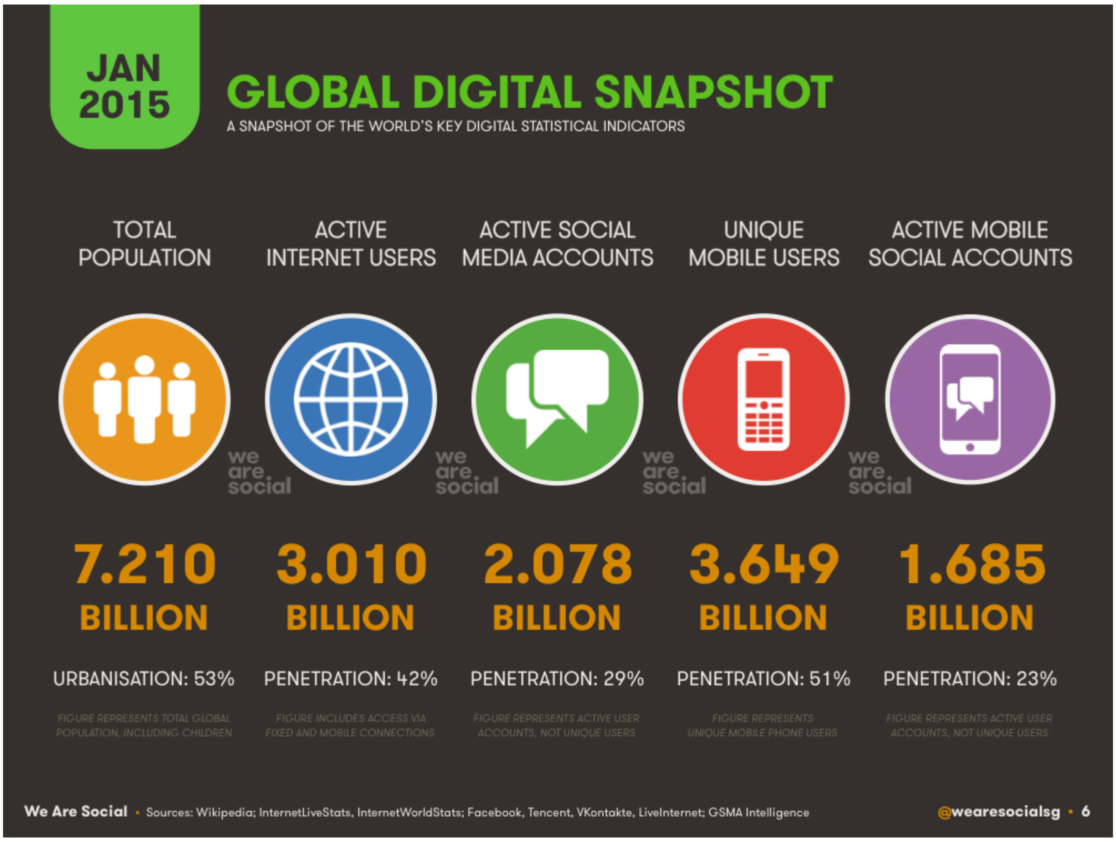
Identification of mentioned cities The entity extraction techniques are applied on the posts in each platform. The posts mentioning place names are identified and the focus point was chosen as 4 big cities;

- Milan,
- Paris,
- London,
- Rome.

The same process of sentiment analysis has been applied on the posts mentioning these 4 cities. The general tendency of polarity and subjectivity the sentiments were analyzed.

4.2 Motivation

There are many studies concentrating on the analysis in different platforms focusing on a variety of aspects of these platforms. But there aren't any work focusing on the relation of posting behavior between different platforms. This is one of the main motivations for choosing the topic, to analyze the relation between different platforms and provide an entry point for future works about this subject. To explain the reason for choosing social network data as a source for the analysis, it is useful to discuss the social networks. ”We are social” (wearesocial.net) is a global conversation agency that helps brands to listen, understand and engage in conversations in social media. A research on the global social media connections has been conducted. [28] The research gives an extensive idea about the global image of social network usage. This research encloses that the 42% of the global population are active internet users (Figure 4.1), and the 29% are also active social networks users. This may seem as a small percentage, but it is essential to recall the fact that the 52% of the global population lives in the urban area.



JAN 2015

GLOBAL DIGITAL SNAPSHOT

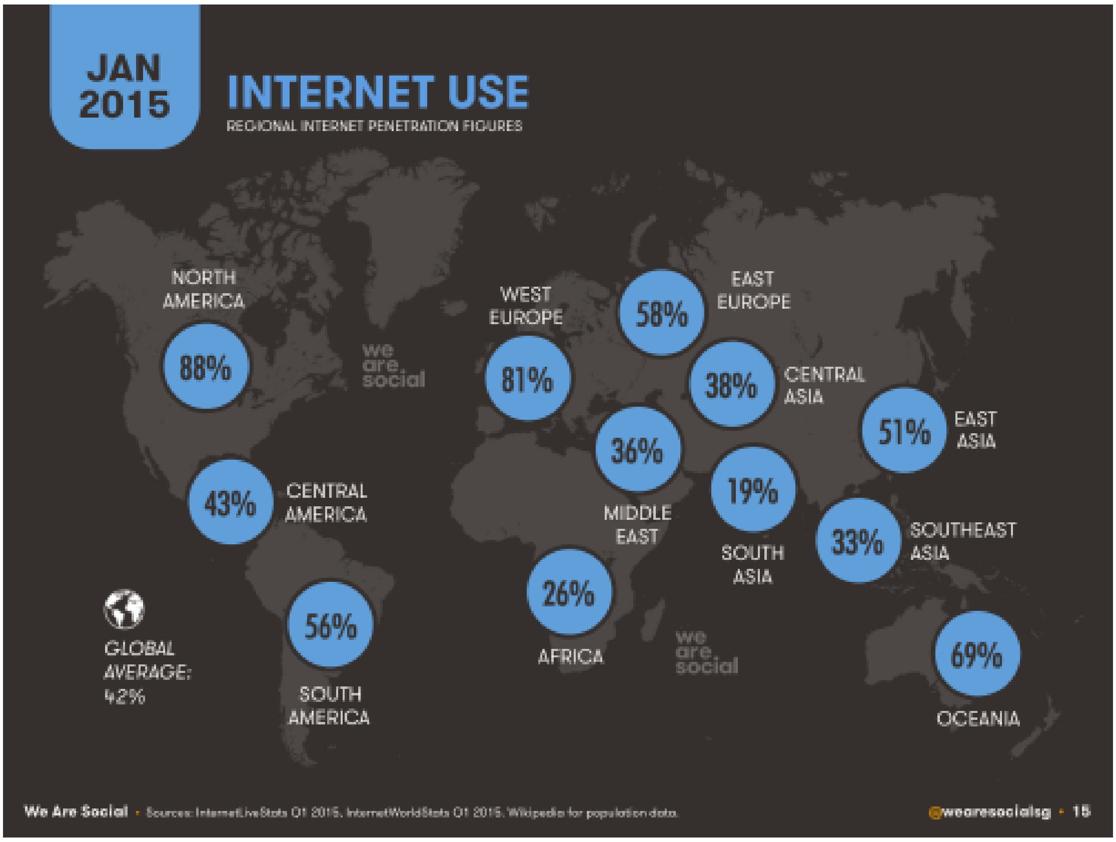
A SNAPSHOT OF THE WORLD'S KEY DIGITAL STATISTICAL INDICATORS

TOTAL POPULATION	ACTIVE INTERNET USERS	ACTIVE SOCIAL MEDIA ACCOUNTS	UNIQUE MOBILE USERS	ACTIVE MOBILE SOCIAL ACCOUNTS
7.210 BILLION	3.010 BILLION	2.078 BILLION	3.649 BILLION	1.685 BILLION
URBANISATION: 53%	PENETRATION: 42%	PENETRATION: 29%	PENETRATION: 51%	PENETRATION: 23%
FIGURE REPRESENTS TOTAL GLOBAL POPULATION, INCLUDING CHILDREN	FIGURE INCLUDES ACCESS VIA FIXED AND MOBILE CONNECTIONS	FIGURE REPRESENTS ACTIVE USER ACCOUNTS, NOT UNIQUE USERS	FIGURE REPRESENTS UNIQUE MOBILE PHONE USERS	FIGURE REPRESENTS ACTIVE USER ACCOUNTS, NOT UNIQUE USERS

We Are Social • Sources: Wikipedia; InternetLiveStats, InternetWorldStats; Facebook, Tencent, VKontakte, LiveInternet; GSMA Intelligence @wearesocialsg • 6

Figure 4.1: Global data snapshot

From the Figure 4.2 it is possible to see that the internet usage exceeds the 20% almost every region of the world. In addition to this fact to get a more extensive idea about the expansion of social network usage through the world we might have a look in Figure 4.3, in the past 12 months internet usage is increased 21% globally, and the active social network users increased by 12%. This fact gives a good reason for concentrating the analysis on social network data.



JAN 2015

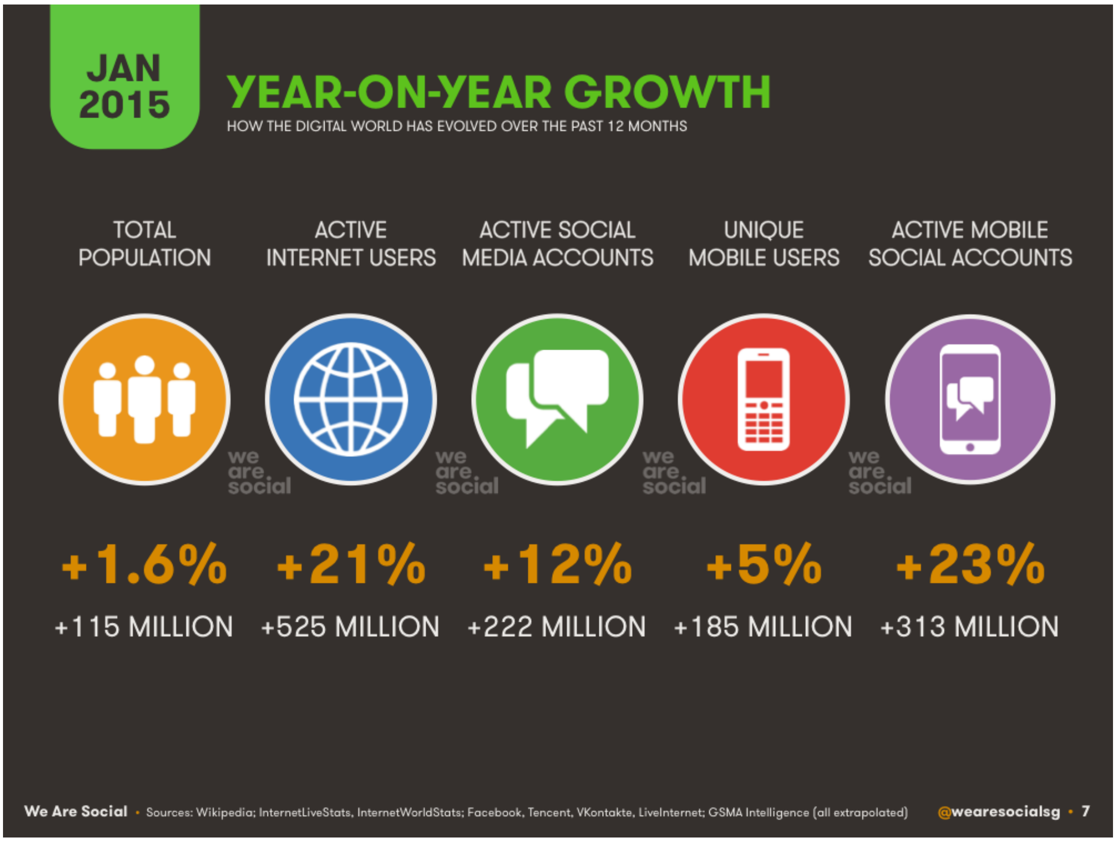
INTERNET USE

REGIONAL INTERNET PENETRATION FIGURES

We Are Social • Sources: InternetLiveStats Q1 2015, InternetWorldStats Q1 2015, Wikipedia for population data.

@wearesocialsg • 15

Figure 4.2: Global data snapshot



JAN 2015

YEAR-ON-YEAR GROWTH

HOW THE DIGITAL WORLD HAS EVOLVED OVER THE PAST 12 MONTHS

TOTAL POPULATION	ACTIVE INTERNET USERS	ACTIVE SOCIAL MEDIA ACCOUNTS	UNIQUE MOBILE USERS	ACTIVE MOBILE SOCIAL ACCOUNTS
+1.6%	+21%	+12%	+5%	+23%
+115 MILLION	+525 MILLION	+222 MILLION	+185 MILLION	+313 MILLION

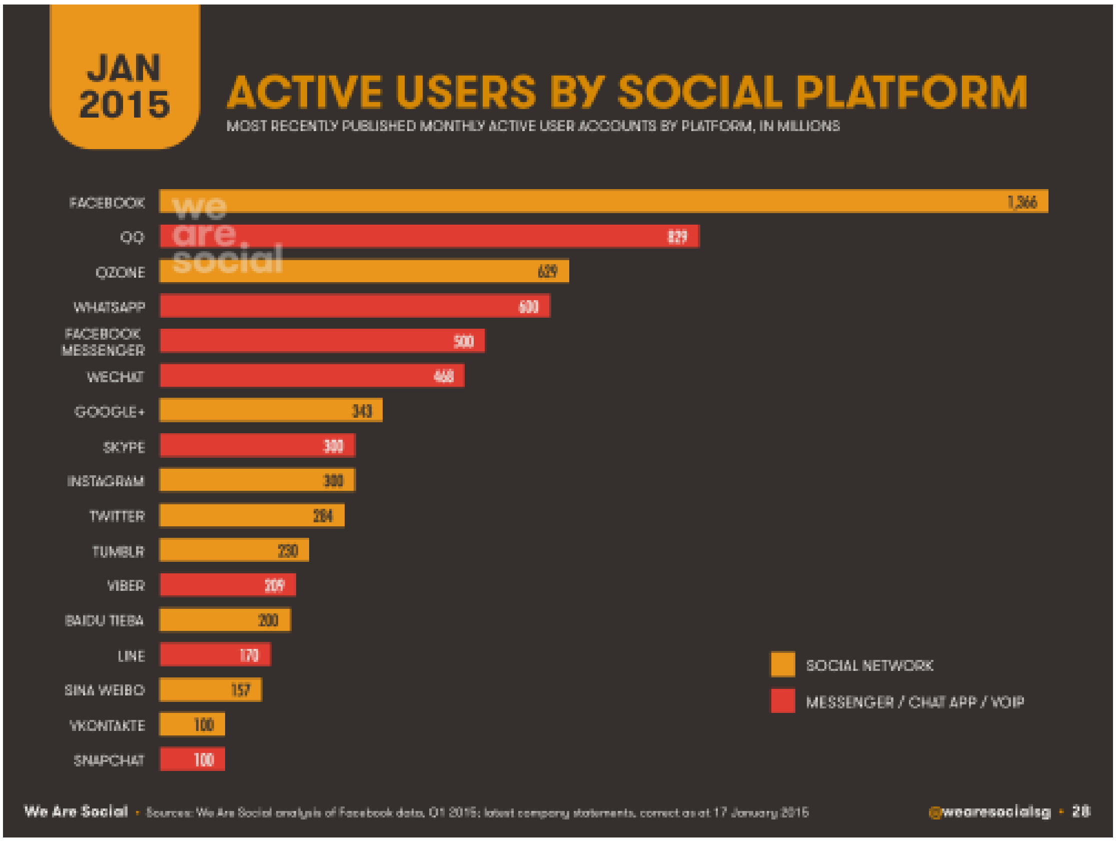
We Are Social • Sources: Wikipedia; InternetLiveStats, InternetWorldStats; Facebook, Tencent, VKontakte, LiveInternet; GSMA Intelligence (all extrapolated) @wearesocialsg • 7

Figure 4.3: Global data snapshot

Focusing on the active users of the social network in figure 4.4 it is possible to recognize that Facebook is the most popular social network through the all. Continuing examining through the most popular social network platforms it is possible to see the names Google+, Twitter, Instagram, Skype, it is useful to ignore QQ and Qzone since they are specially popular just in China. The decision was to continue with Twitter and Instagram from these social network platforms since they provide publicly available information that could be used, on the other side Google plus wasn't preferred since the publicly available data is very limited, the information and posts are usually privately shared by users.

There are few main issues while working with the social network posts,

- They are not categorized, in other words they don‘t provide a precise idea about the contents of the posts, in terms of related topics,
- The insights of the texts in terms of positivity/ negativity are not known; in other words they don't provide a predefined label that can be automatically processed, which gives an idea of the sentiments.
- The entities mentioned in the posts in terms of places, people and so on are not classified.



Sources: We Are Social analysis of Facebook data, Q1 2015; latest company statements, correct as at 17 January 2015

FIGURE 4.4: Global data snapshot

During the process of the relation analysis between 2 different social network platforms, a system to overcome these issues was created. This system gets the idea of the content in terms of topics for each post and puts a label for sentiments, and identifies the mentions for 4 different cities.

4.3 Social Data

This section is dedicated to give an extensive explanation about the data source that has been studied. A conceptual model (Figure 4.5) has been provided, giving the idea of possible ways to search for the correlation. As stated before, the two different social network platforms that have been selected as the focus of analysis: Instagram and Twitter. The structure of the collections will be provided and the possible entities and relationships in these platforms will be discussed:

- Instagram : Instagram is the main platform for sharing photos. In this platform, the photos are shared by the users with an optional additional caption, these captions give an idea about the content and sentiment about the photos that are shared. The followers can make a comment or like on the post, mention each other, the users can add geolocation to their photos before posting. The hashtags are vastly used in the captions to give the subjective idea of the user, or to describe the picture from the poster's point of view.

- Twitter: Is a more generic type of platform, a variety of information can be obtained, but usually it is relevant for textual data. Messages are identified by the entity Post, that are written by an author and can be or not geolocalized in a Place. In a message, there are many kinds of information that can be extracted. First of all, the text is referring to a topic, that can be identified with the entity Category. In addition there are Hashtags, keywords that can be inside a message, and can usually give the main idea of the content. Second, in a text can be present mentions of a Place, or a Person.

For the posts in both platforms exists the entity Person, that identifies the author of the contents published, and a relation between different persons. For every social network post, there is the text, and the born-time, which is used extensively for analysis through the time interval. In addition to the text for each post, there are entities related to this post including hashtags. Places, Links, people that are mentioned in the post, are labeled as entity generically and stored with the text. In the preprocessing stage of the work the entities didn‘t exist in the structure, after the processing and extraction, these new data are appended to the post entry. The same process is executed for the sentiments.

It is useful to state that in Instagram there is the optional caption which includes the text related to the photo and the array of hashtag entities inserted by the user. In cases where the text of the caption didn‘t exist, the text of the hashtags are concatenated to each other are used as the text field to give a unified view. These hashtags and text are used in the following steps for content analysis, extraction of sentiments and extraction of place names.

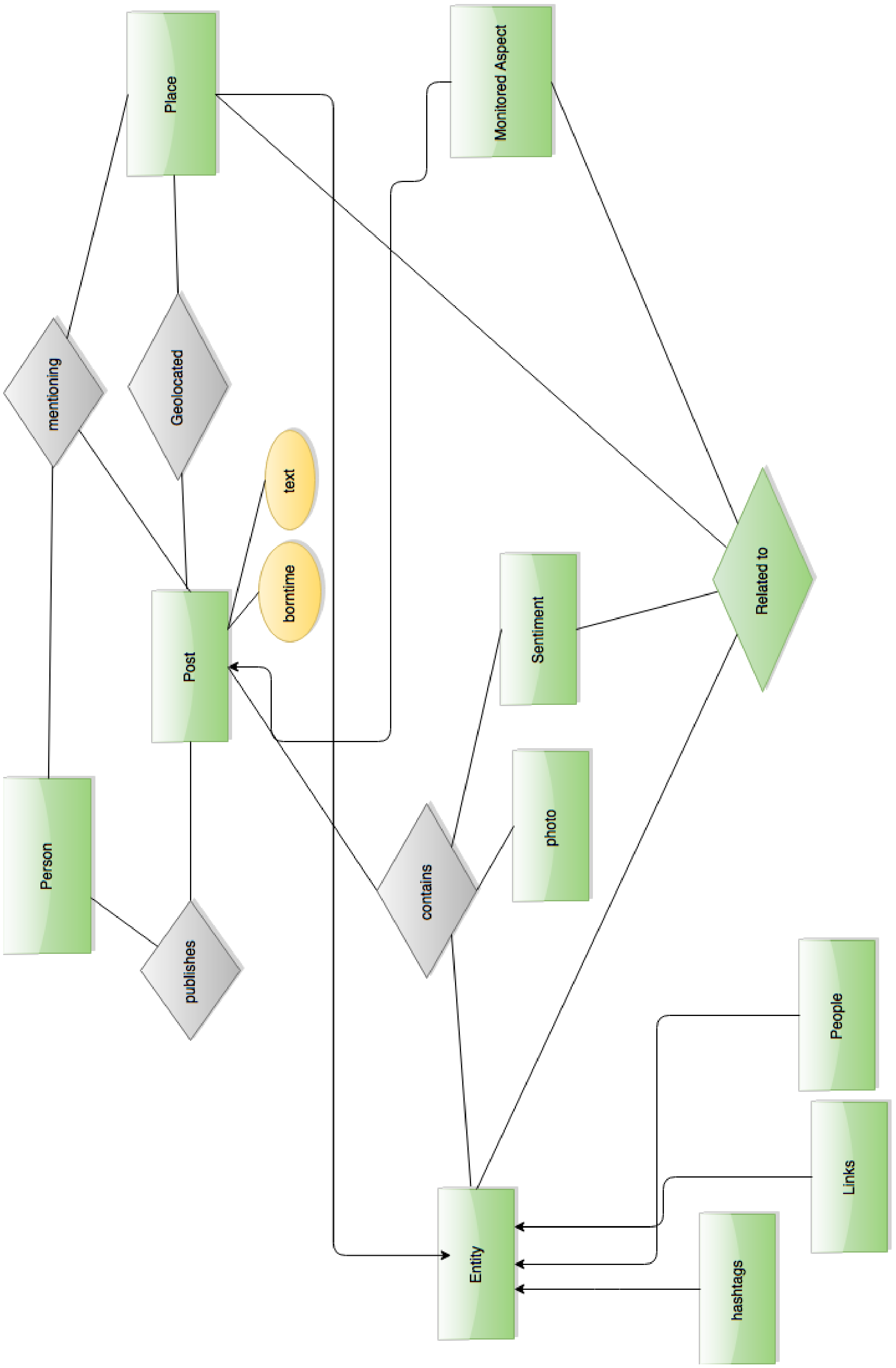


FIGURE 4.5: Conceptual schema

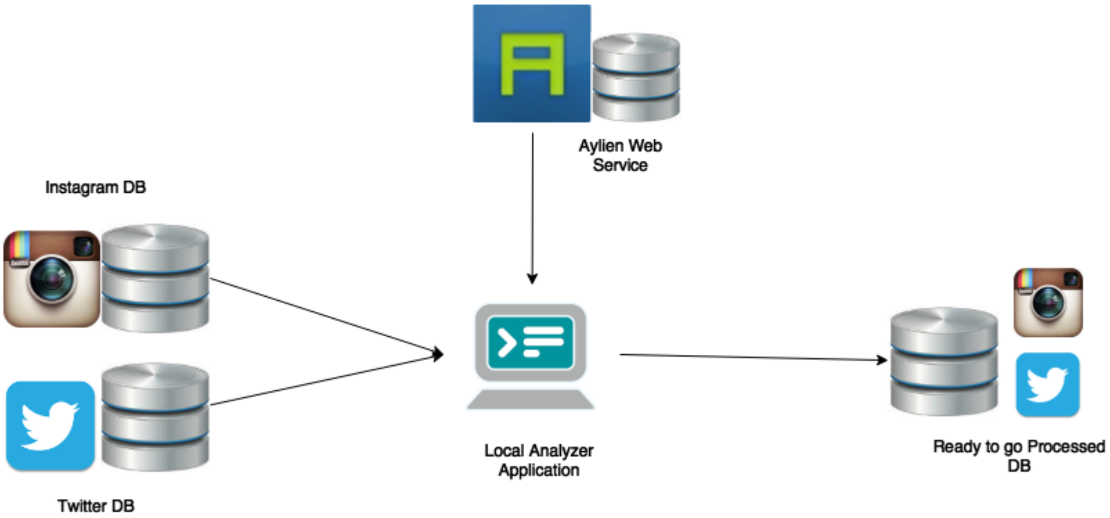


FIGURE 4.6: Architecture Design

4.4 Approach

The approach proposed, consisted of many steps following each other, to draw a main picture, these steps can be generalized into 3 main phases that have been followed during the whole process :

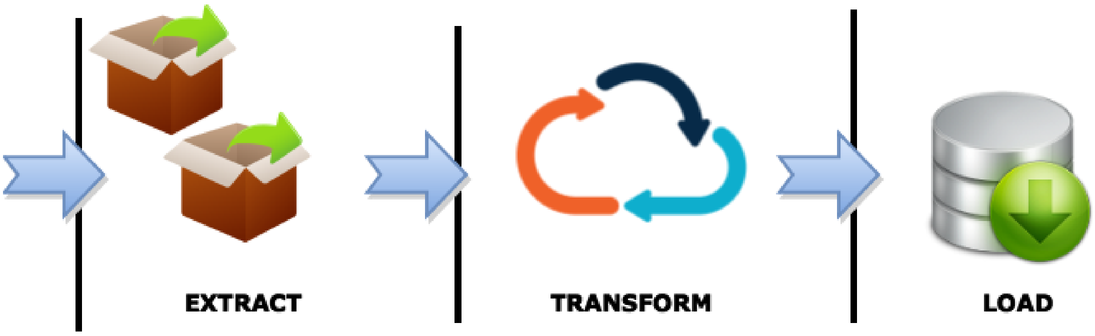


FIGURE 4.7: Phases

The unprocessed data is consisted of two collections in different formats covering different time intervals. In order to put the data in a form that will be appropriate and adequate for the correlation analysis, the well known ETL paradigm is chosen to be applied:

- Extraction Phase
- Transform Phase:

  - Language Detection
  - Sentiment Extraction
  - Entity Extraction

- Load Phase

These phases are followed by experiments performed on this data as well as the representation and discussion of the results.

4.4.1 Extract Phase

This is the phase where the raw data is extracted. The unprocessed data analyzed in this phase were in different formats. In other words, it was heterogenous, and covering different time intervals.

According to these facts about the raw data, 2 things were needed to be done :

- Identification of the interval that is covered by both datasets and the extraction of the posts that has the born-time in this intersection interval.
- Mapping the different formats between these two collection to obtain a simple format which could be used for both social network posts, for the sake of simplicity in the experiments part.

For achieving these main goals in the extraction phase, the querying had been done, in terms of

- the created date, if the post corresponds to the identified time interval to be extracted.
- the content, to extract the posts in content to be analyzed, changes depending on the analysis.

4.4.2 Transform Phase

The transform phase of the process performs all the processing part of the raw data, to be appended to the new dataset which will be used for analysis and experiments. This phase consists of 3 main steps to be followed.

4.4.2.1 Language Detection Phase

For the following steps of the transform phase it was essential to know the language of the text of each post that are extracted. For this reason, the language detection phase needed to be the first step to be taken was the language detection.

In Twitter collection the language field has already existed, on the other side in Instagram dataset the entries didn't have the field specifying the language. For identification of the language of the Instagram entries the java library langDetect[2] is used. After the detection of the language, for each Instagram post, the a new language field is added to the posts to be used in the following steps.

4.4.2.2 Sentiment Extraction phase

In order to search for correlation between the sentiments of different platforms in the experimental process, the sentiments for each extracted post needed to be identified. To achieve this goal, the sentiment analysis functionality of Aylien text analysis API [3] is used.

During this process for each extracted post, the ones having the language English as a label are selected, since the sentiment analysis functionality is available only for the English language. This part is achieved by querying the extracted dataset by language. In the next step, the text contained in the post was extracted and the web service was called to identify the sentiment of the text. It is useful to state during explaining this process that, all Twitter posts included a text field, on the other side the for posts in Instagram there were some entries that didn't contain any text but just hashtags. In this case the hashtag texts are appended to each other and were sent as a parameter to the web service for identifying the sentiments. The response from the web service contained a data structure with the fields about polarity, subjectivity and the confidence values ranging from 0 to 1 for both of subjectivity and polarity labels returned. This data structure is appended to each post entry for the following steps.

4.4.2.3 Entity Extraction phase

In order to make the analysis to search for correlation between the sentiments of the posts concerning specific entities in different platforms, the entities mentioned in each extracted post needed to be identified. To achieve this goal, the entity extraction functionality of Aylien text analysis API [3] is used.

During this process for each extracted post, the ones having the language English as a label are selected since the entity extraction functionality is available only for the English language. This part is achieved by querying the extracted dataset by language. In the next step, the text contained in the post was extracted and the web service was called to extract entities from the text. Just as the sentiment extraction phase, for posts in Instagram that didn't contain any text but just hashtags, the hashtag texts were appended to each other and sent as a parameter to the web service for extraction of entities. The response from the web service contained a data structure containing many possible lists; organizations , locations, persons, notable keywords, date values, money values, percentage values, time values, URLs, email addresses, phone numbers that are found in the text. In case of no values found, corresponding to one category, that list doesn't exist in the response. This data structure is appended to each post entry for the following steps. As we are interested in the locations, the posts with the entities having the location list not empty will be analyzed in the following steps.

4.4.3 Load phase

In this phase following the processing of the data, the storage of data is performed, data has been given a more basic and standardized structured format, cleaned from the unnecessary fields (see data design in the next chapter). This new structure combined with the extracted information (sentiment and entities) is stored in the new dataset, for serving the simplicity of following steps.

4.4.4 Observation and Analysis Phase

This is the phase where the new dataset is used for making new observations and analysis. These observations consist of many different experiences each focusing on a different aspect of the data for the possibility of enclosing a new useful information about the data source.

This subprocess consists of execution of many different queries on the new dataset, the evaluation of the results in terms of correlation coefficients, percentages or shared numbers, as well as the graphical representation and interpretation of the results.

Chapter 5

Implementation Experiences

To validate the proposed solution for the analysis of the correlation of contents, shared in Instagram and Twitter, a system has been developed to cover all the phases explained in the Chapter 4. This chapter is aimed to show and explain the steps which led to the development of the system that is used in the experimental part, in terms of technical aspects.

5.1 Method

As the thesis work is mainly a research project, there isn't any predefined solution and a method to follow step by step to realize the proposed solution. For this reason the Spiral Model for the code implementation, which provides a model for developing a system through a sequence of incremental versions is used.(Figure 5.1).

This model was first described by Barry Boehm in his 1986 paper "A Spiral Model of Software Development and Enhancement" [29]. The main characteristic of this model is its feature of being cyclic, differing from other models which are linear, as waterfall model and incremental model. In this model each cycle is composed by 4 phases:

- identification of goals and requirements,
- identification of the risks,
- development of the prototype,
- review of the results.

The model proposes the software to be developed in a sequence of increasing versions. This was very suitable to apply on the solution because for every step of analysis, there were different requirements to be covered and each prototype should be developed to satisfy the corresponding requirements.

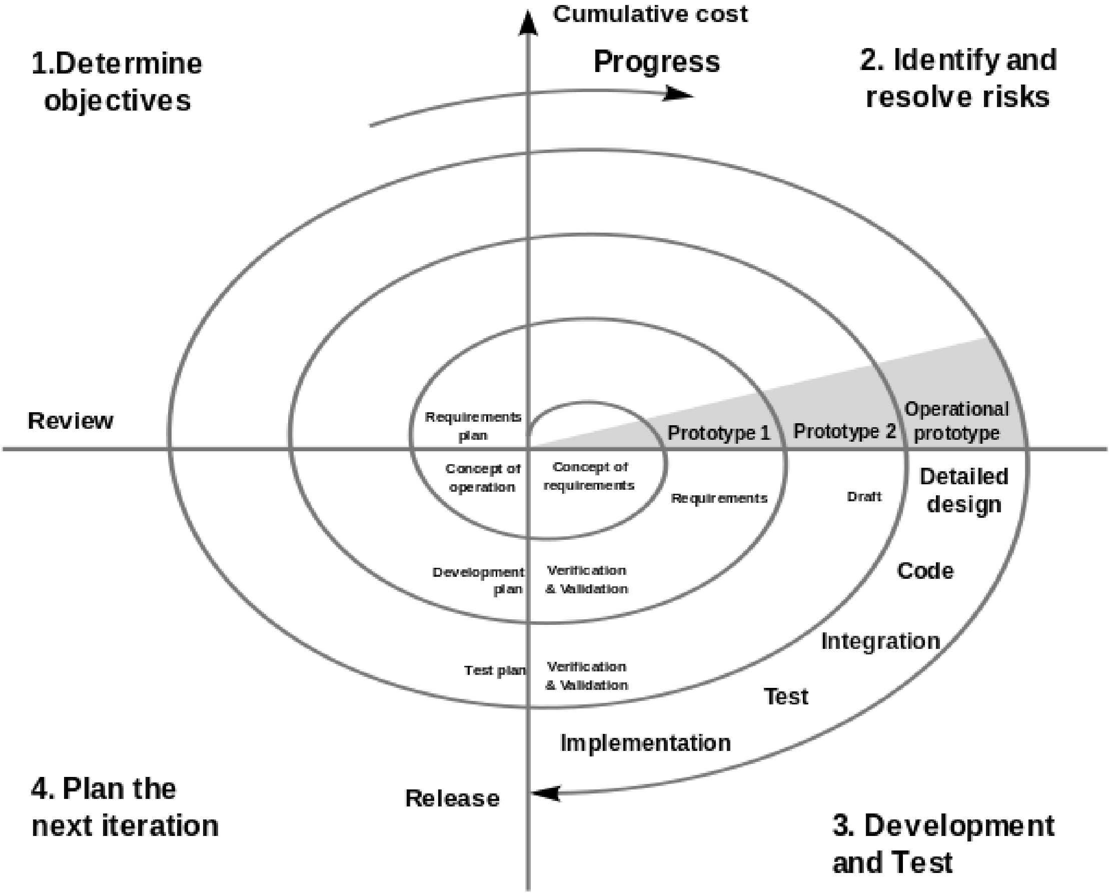


FIGURE 5.1: Spiral Model

The main advantages of using spiral model are:

- The development phases can be determined by the developer according to their complexity which makes it one of the most flexible models,
- Project monitoring is very easy and effective. Each phase and loop, requires a review and approval. This makes the model transparent in terms of validation for each phase,
- The risks are needed to be identified and managed for each single phase of the spiral. This makes this model more attractive than the other ones in terms of risk management,
- The deadlines can be better managed in terms of developed code and reached goals;

- It is a very suitable model for highly customized projects.

The main disadvantages of using spiral model are

- The development costs are generally higher than the others,
- For the projects with clear and predefined scopes, it is an unnecessary approach because of its complexity. For these projects, linear models would provide an easier and faster solution,
- The approval and validation phases require a review of an expert,
- The highly customized structure makes the prototype very hard to reuse for other projects.
- The model requires a high amount of documentation for intermediate stages which increases the complexity of the development process.

The programming language java is chosen for the implementation of the prototype because of its flexibility, easiness, and available libraries. Eclipse is chosen as a development platform because of the familiarity. The first prototype was implemented to realize the extraction phase, in other words, the system for extracting the entries one by one from both datasets covering the predefined interval. In the following steps new functionalities are added to this prototype incrementally. First, the system for hashtag analyzes is implemented and added to the project, then after the validation and visualization of the results, the language detection part is added. Following this step the system for realizing sentiment extraction and entity extraction has been implemented, the new datasets are created. Eventually, the part for analyzing these additional data has been added to the prototype to give the complete picture.

- In the implementation of extraction part which queries data from Mongodb datasets, for the simplicity of the querying, mongo java driver as well as Jongo has been used. "MongoQueryManager" class was created to produce the different queries according to the tasks and parameters to serve for the extraction process.
- For implementing the subsystem that accomplishes the hashtag analyzes, "MongoDBMgr" class was created. The functionality implemented in this class for the hashtag analyzes needed to perform weekly monthly and daily extractions for the selected time interval. The JodaTime library is used for spanning the weeks days or months of the interval. Each entry is first cleaned from stop words and platform specific tags, then the mappings had to be done according to predefined

rules. Following this essential step that should be performed for each entry having a timestamp in the given time interval, the grouping was done according to the hashtags.This goal is achieved by using mapreduce functionality of Mongodb. "MongoMapReduceQM" class was created to manage and produce the different map and reduce functions according to given parameters, to be executed on the datasets. The map function for this part does the cleaning and mapping part and emit each tag of the each entry in the given time interval. In the reduce function, each different hashtags and numbers of the occurrences corresponding to these tags are calculated.

- For implementing the system that detects the language of the entries a Language Detector class with a functionality that uses the LangDetect [2] library for guessing the language of the inputted text is implemented. This functionality returns an array of languages with the probabilities corresponding to each element. For each entry extracted from the dataset, the language array is evaluated. In this stage of the process it is important to specify that especially in instagram dataset there are many entries that contain text in multiple languages. To use as many entries as possible for the following extraction step, an entry is labeled as English if there is a probability higher than 50%.

- In the implementation of the subsystem that extracts the sentiments and entities for each entry, Aylien[3] extractor class is created. This class contains functionalities that given a text input, makes the web service calls and returns the extracted entities or sentiment. In the mongodb Manager class for each entry these extraction functions are called. The new information containing the sentiment or entities is appended to each post and this new structure is appended to the post and stored in a new dataset.

- For the implementation of the subsystems that accomplishes the analysis other functions are implemented in the "MapreduceQueryManager". One of this functionalities is for evaluation of the sentiments of the most popular hashtags for some time interval. For achieving this goal, the map function, after accomplishing cleaning steps in the same way as explained in the previous point, emits the hashtag as well as the sentiment. In the reduce function for each tag, the number of occurrences is calculated and the number of entries with positive, negative, neutral, subjective and objective entries calculated separately.

5.2 Software Design

The system implemented is a local application which processes the raw data, uses web services for performing information extraction and creates a new dataset with additional values.

These values are used by the same application in the analysis step. The architecture representation can be seen in Figure 5.2.

The implementation consists of 2 packages, DB, and Utility. In Figure 5.3 the structure of the packages and classes and the interaction between them can be visualized.

- DB: Contains the classes for handling the database connections, queries, as well as the utilities and structures that are used for managing and manipulating the data.
- Utilities : Contains the classes for detection extraction as well as the structures for manipulation of the results

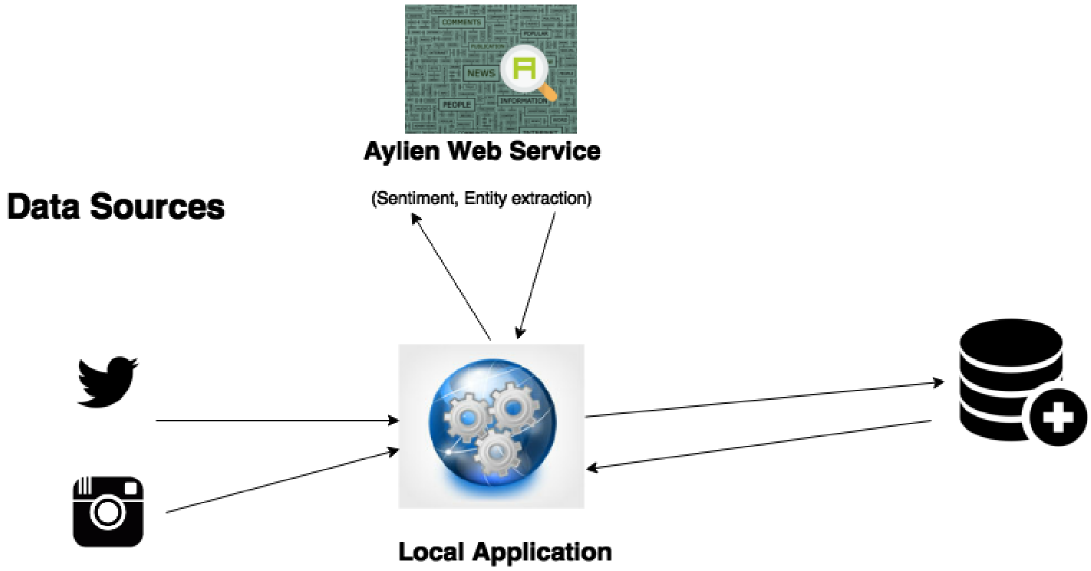


Figure 5.2: Design of the Architecture

Here is the list and functionality of the classes:

- MongoDBMgr : This class as can be understood from the name, deal with initiating any functionality related with the databases. Contains methods that do loading, saving, querying, processing of the required data.

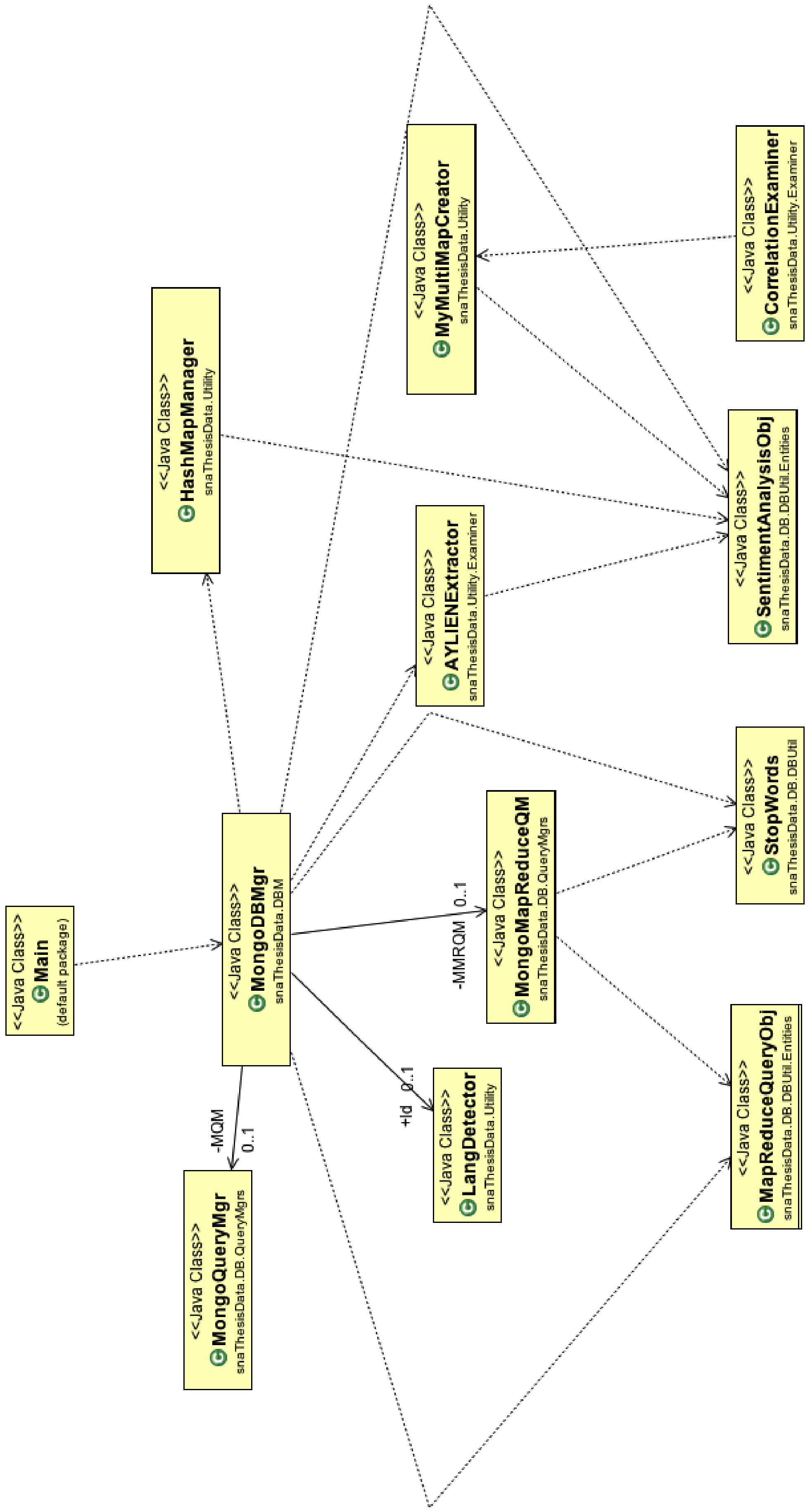


Figure 5.3: Class Diagram

- MongoQueryMgr: This class serves as a query production tool, to be used by MongoDBMgr, according to given parameters and called methods the class creates queries and returns them in required formats.
- MongoMapReduceQM: The class contains functionalities that creates Map and reduce functions to be executed on datasets for different aims. This is done according to called methods and given parameters.
- MapReduceQueryObj: This object is used by mongoMapReduceQM, gives a structure for the methods returning map and reduce at the same time. The objects have the map and reduce fields which are set by the MongoMapReduceQM functions.
- MultimapCreator: Used by other classes for aggregating 2 different hashmaps in different formats into a hashmap. In some cases, this is done for creating a hashmap with combined values to be used for correlation calculation.
- HashMapManager: Used by other classes for manipulation and processing of hashmaps, this class provides functionalities as, sorting hashmaps, transforming them to a list, printing hashmaps.
- AylieneExtractor : The class as can be understood from the name, uses java SDK for Aylien to accomplish the entity extraction as well as sentiment extraction.
- CorrelationExaminer: The class uses apache commons Math libraries' statistics package to calculate the correlation coefficients for 2 arrays inputted.
- LangDetector: The class uses the LangDetect library to identify the language of the inputted text, and return the array of languages with corresponding probabilities or the string of the language code.
- StopWords : Contains the stopwords and platform specific tags to be cleaned in the analysis process.
- SentimentAnalysisObj: Provide a structure for storing the tag string, number of positive, negative and neutral occurrences as well as the number of subjective and objective mentions.

5.3 Software Flow

The software works in the following way: The main class initiates the MongoDBMgr and the functionality for extraction of entries that corresponds to the intended time interval. For every entry in this set returned from the query, langDetect class' detection functionality is called. After the language is labeled and added to the dataset, a new functionality is called for extraction of sentiments and entities. After the new dataset is created, the methods for analyzing the new data from different aspects, are called and the results are visualized. (Figure 5.4)

5.4 Data Design

In this section, it is planned to explain the structure of the data that is stored in new datasets, to be analyzed. The new datasets are designed to contain all the useful information in the most efficient way as possible. To represent the MongoDB collection data structure, Atlanmod JSON Discoverer is used.[30]

The first structure is designed for the Instagram posts. This structure consists of many substructures:

- User: The sub-collection that represents the author of the post. This sub-collection contains fields for
  - Id: The unique identification number for each user
  - Screen_Name: Screen name of the user
  - Profile picture: URL Link to user's profile picture
  - Created_date: Timestamp responding to date of creation of the profile
- Caption: The caption inserted by the user which contains an explanation about the photo.
- Comments: Array of sub-collection of data, which contains
  - id: The unique identification number for the comment
  - text: Text in the comment.
  - created_time: Timestamp for the created time.
  - userName: Name of the author of the comment.
- Location: The location of the post, if it is geotagged. Contains coordinates of the location.
- Language: The language code of the post, that is extracted.
- Likes: Array of sub-collection of data, which contains the number count,
  - id: The unique identification number for the comment

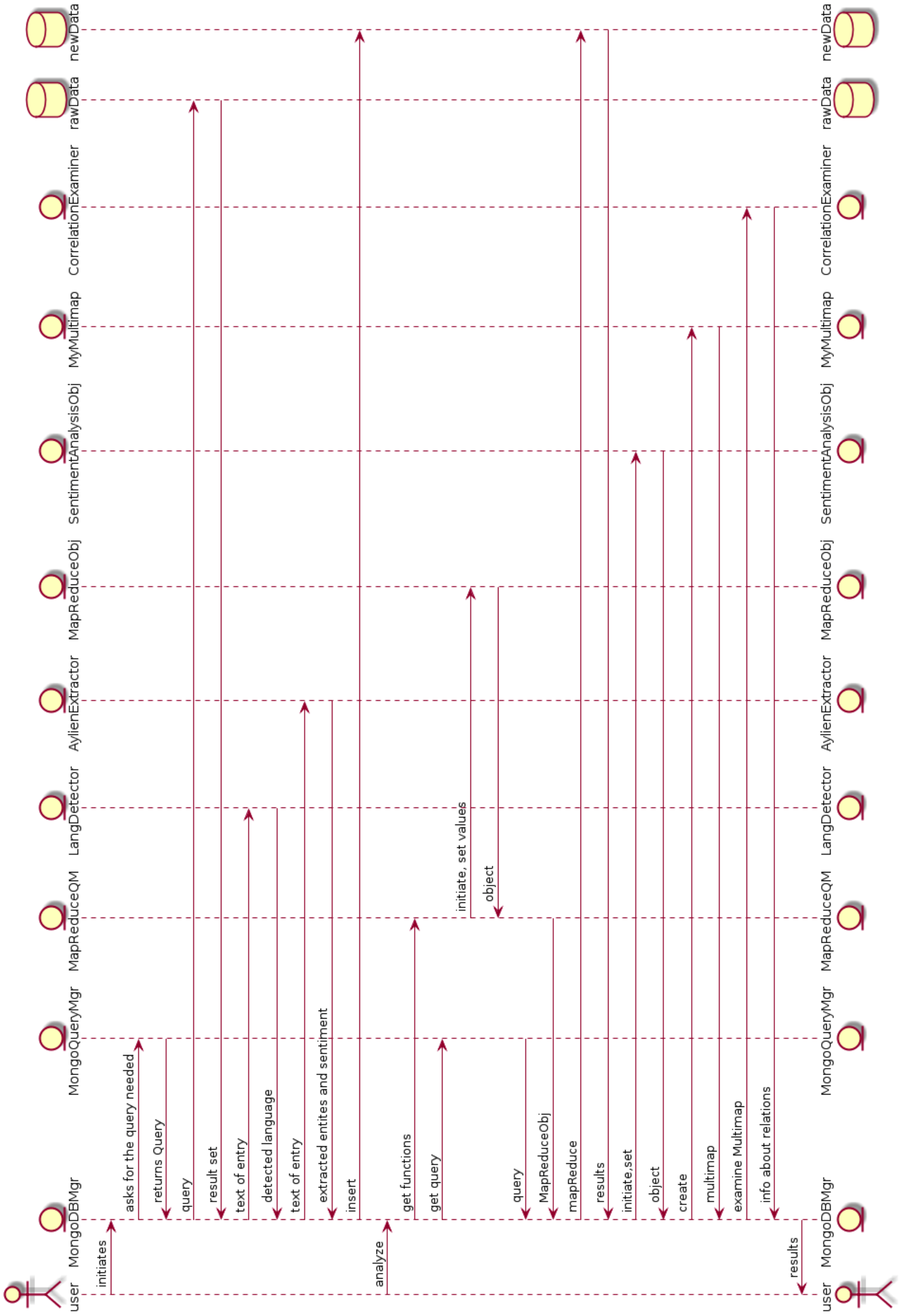


Figure 5.4: Sequence Diagram

  - text: Text in the comment.
  - created_time: Timestamp for the created time
  - userName: Name of the author of the comment.
- Tags : Array of hashtags which are contained in the text.
- Entities : Array of entities that are extracted from the text, can contain array of locations, people, emails, brands, percentages, phone numbers, keywords...
- Sentiment : The sentiment that is extracted from the text, includes a label for polarity and a label for subjectivity, with the number for confidence changing from 0 to 1.
- Images: Contains different images according to resolutions, for each image exists
  - URL String linking to the image
  - width of the image
  - height of the image

Another structure is the one representing tweets. This structure contains following fields:

- Text : The full text of the tweet
- Language :Language code that is detected
- Author : Name of the author
- createdDate : Date of creation as a timestamp and following sub-collections:
- Entities : Array of entities that are extracted from the text, can contain array of locations, people, links, mentions, emails, brands, percentages, phone numbers, keywords..
- Sentiment : The sentiment that is extracted from the text, includes a label for polarity and a label for subjectivity, with the number for confidence changing from 0 to 1.
- Location: The location of the post, if it is geotagged. Contains coordinates of the location.
- Raw: A sub-collection that is contained in the raw dataset for each entry. This sub-collection although is not used in analysis part, it is kept for the further work. Includes;
  - id: The unique number for identifying the post.

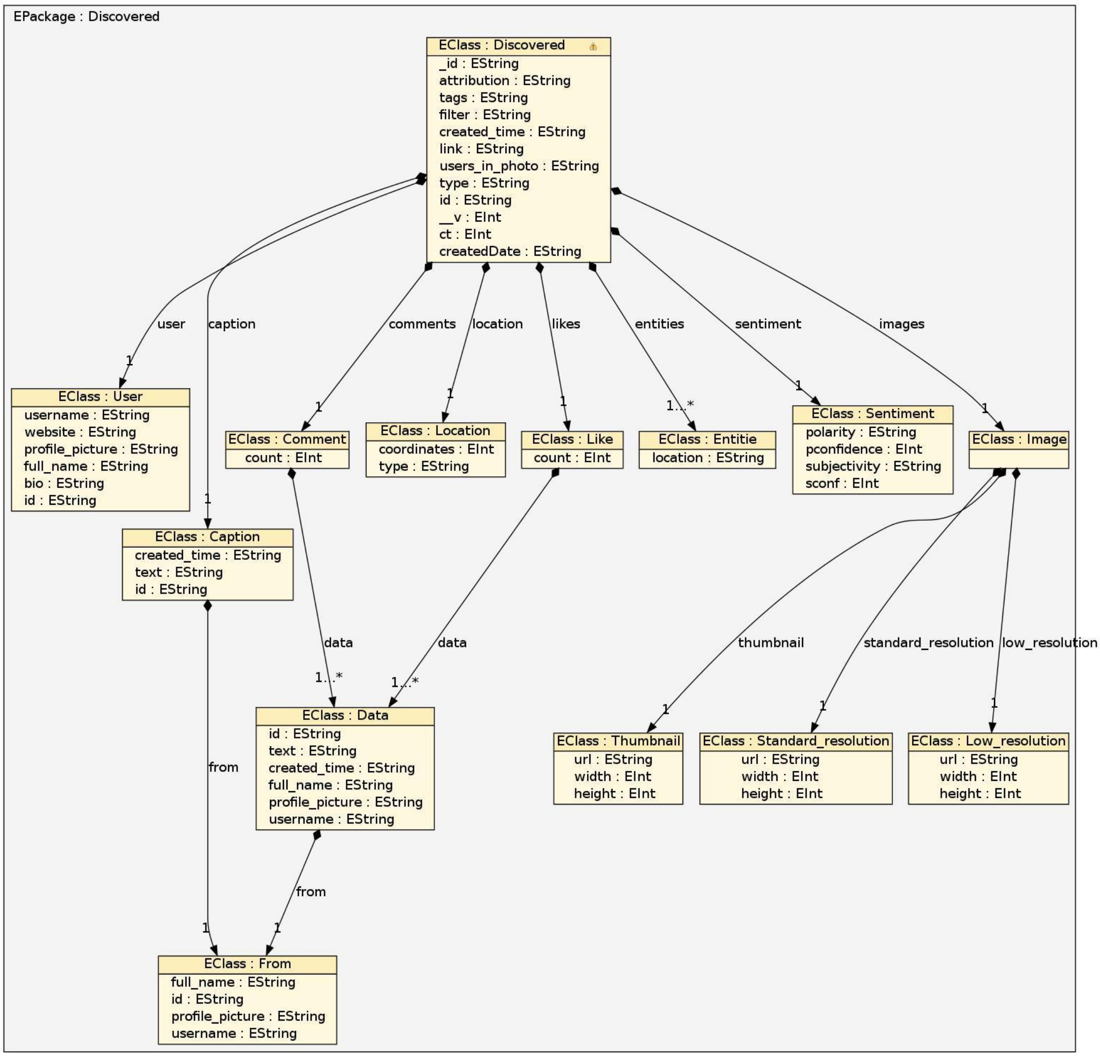


Figure 5.5: Figure 5.5: Instagram post data structure

- FavoriteCount: the number of favorites
- RetweetCount: the number of retweets
- Retweeted: boolean if the post is a retweet
- In reply to user id: The user id, to whom the post is written in reply to,
- In reply to status id: The status id, to whom the post is written in reply to,
- User: User sub-collection, containing detailed information about the author.

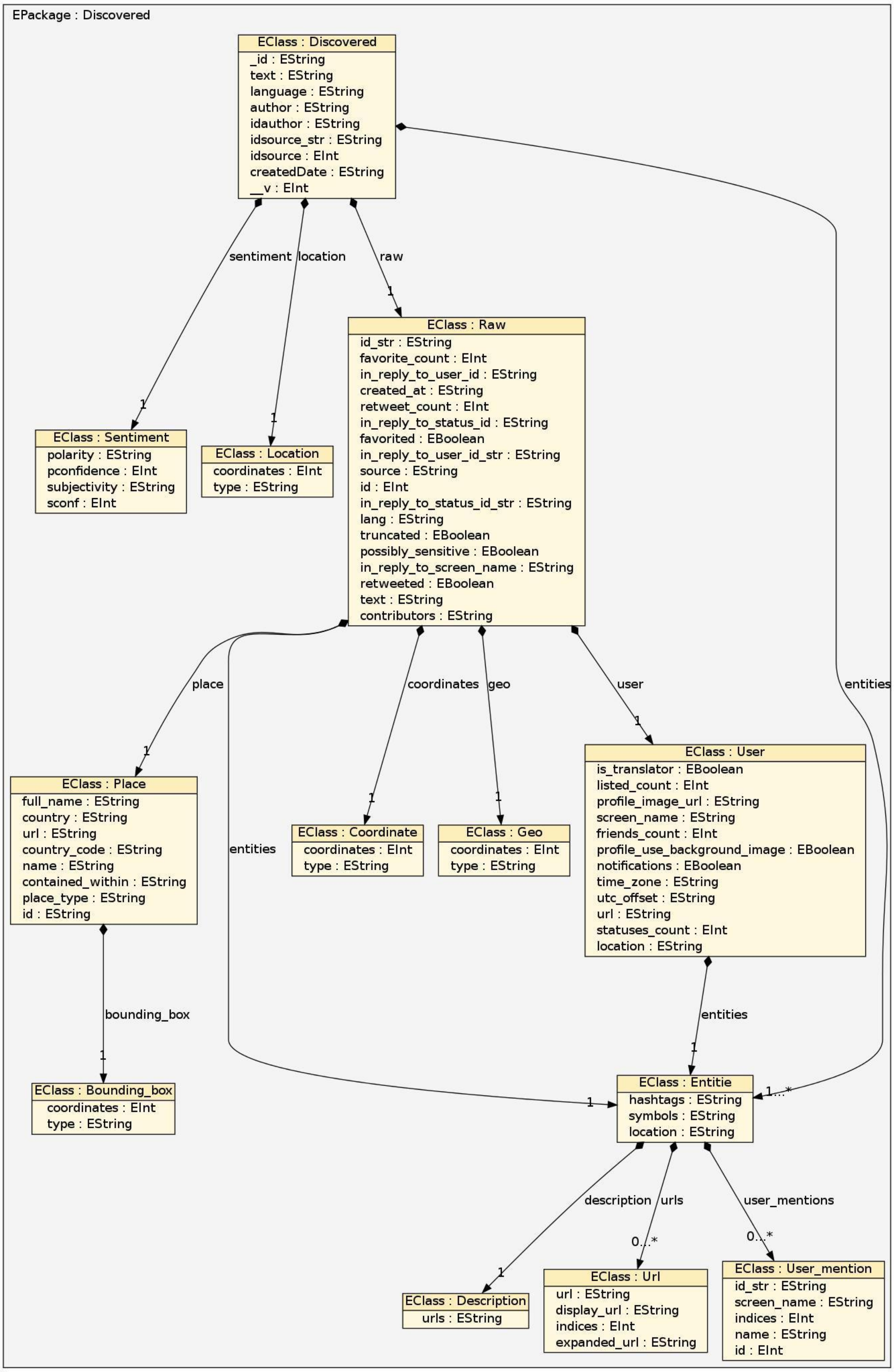


Figure 5.6: Twitter post data structure

Chapter 6

Experiments and Discussion

In this section of the document, the experiments that have been done for specifying if there is a relationship between Twitter and Instagram will be explained one by one. The process and the aspects that have been considered will be indicated, the results will be provided in different forms as graphs, diagrams, the results will be discussed afterwards.

Dataset General Information:

The dataset consisted of two mongodb collections, one for Instagram posts, one for twitter posts that are collected during the process of previous research for analyzing the post relations between the 4 big cities Rome, Milan, Paris, London.(Instagram db from Topics of Interest in Cities Based on Content Mining from Social Networks [9] and Twitter DB from Monitoring Urban Mobility With Social Networks [11]).These datasets contained Tweets and Instagram posts covering different time intervals and having different formats.

The dataset contained 280.540 Instagram entries between dates 03.07.2013 to 11.12.2014 and 306.598 Twitter entries between 30.05.2014 to 27.03.2015. Summing up to more than 500K entries.

The intersection of intervals was from 30.05.2014 to 11.12.2014. And in this interval covered

* 119996 tweets

* 238937 Instagram posts

119996+238937 entries summing up to 358.933 social network posts having been analyzed.

6.1 Analysis 1 Topic Extraction / Correlation Analysis

The first analysis was to search for correlation of the topics that are most mentioned in the two platforms : Before continuing with analysis itself, there were some parameters to decide on:

- Topic extraction technique,
- The method to evaluate correlation strength,
- Granularity in time,
- The sample to choose from the bag of tags to analyze in this interval,
- The data cleaning approach if needed.

6.1.1 Deciding on Topic Extraction Technique

As it is stated before, the NLP tools or web services using NLP tools that are providing topic categorization remain unsatisfactory for labeling the category of the short texts especially the social network posts because of their special content, on the other side since the web service that has been used in the analysis process had the categorization functionality, an experiment has been conducted to see the results in order to decide whether to use the hashtags as in [7] for representation of the content, or the categories by NLP techniques. To achieve this goal, AYLIEN web service's classification functionality on the texts of Twitter database has been used.

Results

text: #Paris #Marketing #Job: EMEA Corporate Communications Director at SunPower http://t.co/4AZn9Jknr7 #VeteranJob #Jobs #TweetMyJobs,

categories: unrest, conflicts and war - crisis

text : Next stop : Paris. Then hello Dubai, Miami & Las Vegas,

categories: hobby - shopping

text : Faut pas insulter wayy,

categories : lifestyle and leisure - adventure text : I'm at Le Pain Quotidien - @lepainquk (London, Greater London) w/ 2 others https://t.co/nSamBtpObG,

categories : unrest, conflicts and war - crisis

text : @Chris_tiee I don't know I was just walking to brick lane with my cousin and her bf and he was just strolling along,

categories : education - teaching and learning

Discussion of Results

Examining the accuracy level of this functionality manually, it was possible to state that it is very low for the twitter posts. For this reason, the hashtags for the representation of the content of a social network post was preferred to be used.

6.1.2 Deciding for granularity

The first analysis was aimed to find out the most trending hashtags for some period and see if they are consistent in the 2 different datasets, in other words, to identify if there is a relation that can be observed between the contents of the two different platforms..

There were different possible granularities to choose; daily, weekly, monthly. The weekly granularity for the analysis was chosen, since it would be more useful considering the trending subjects, and social network dynamics in the time interval. A random interval of 7 days from 11.11.2014 to 19.11.2014 was selected.

There are 1961 entries in Instagram and 2057 entries in Twitter dataset for this interval.

Results

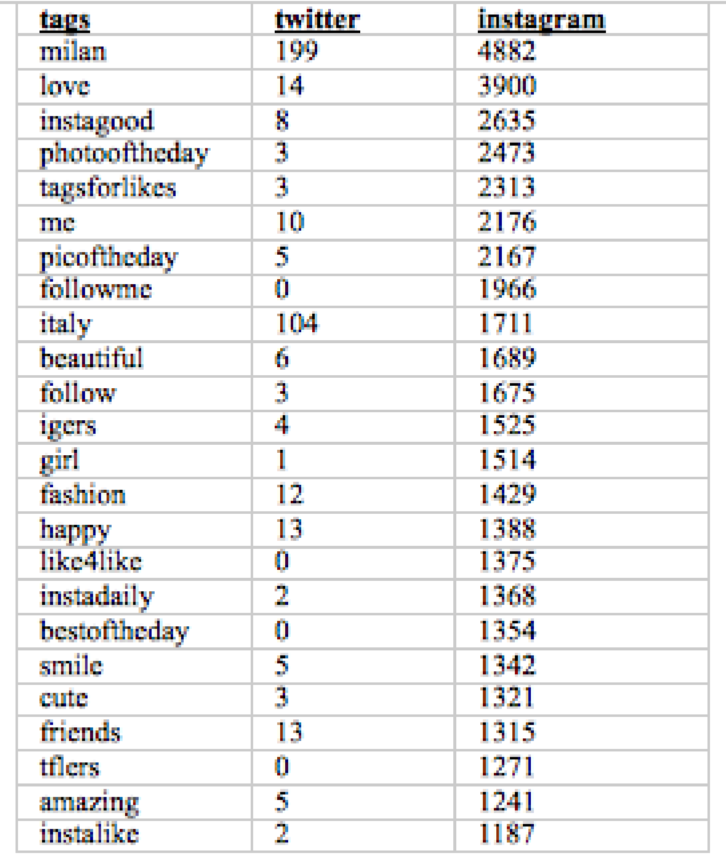
→ Number of hashtag occurrences in Twitter for the interval: 823

→ Number of hashtag occurrences in Instagram for the interval: 13941

→ Hashtag/post in twitter: $\frac{823}{2057}$ =0.4

→ Hashtag/post in Instagram: $\frac{13941}{1961}$ =7.1

The graph of all tags including 5987 different tags that is used by people for the week analyzed:



tags	twitter	instagram
milan	199	4882
love	14	3900
instagood	8	2635
photooftheday	3	2473
tagsforlikes	3	2313
me	10	2176
picoftheday	5	2167
followme	0	1966
italy	104	1711
beautiful	6	1689
follow	3	1675
igers	4	1525
girl	1	1514
fashion	12	1429
happy	13	1388
like4like	0	1375
instadaily	2	1368
bestoftheday	0	1354
smile	5	1342
cute	3	1321
friends	13	1315
tflers	0	1271
amazing	5	1241
instalike	2	1187

Table 6.1: Example of most popular tags (included platform specific tags)

Discussion of Results

After analyzing the bag of tags and the number of occurrences per each platform it was possible to identify 3 important issues.

- Platform Specific Tags: The hashtags which are specific to one of our platforms. Especially there are many hashtags that most probably belong to Instagram. For example instagood, instapic, igersmilano, tflers.. When we check the occurrences of these tags we are able to see that they are high in Instagram and almost do not exist in Twitter.

- The hashtags that give no idea about the topic especially the ones that are used to collect more followers or reach to more users. For example the ones that contain "like", "follow", "tags" etc..

- The tags that refer to the same subject but because of language or syntax mistakes, grouped in different sets, hence giving us the wrong number of occurrence for the topic.

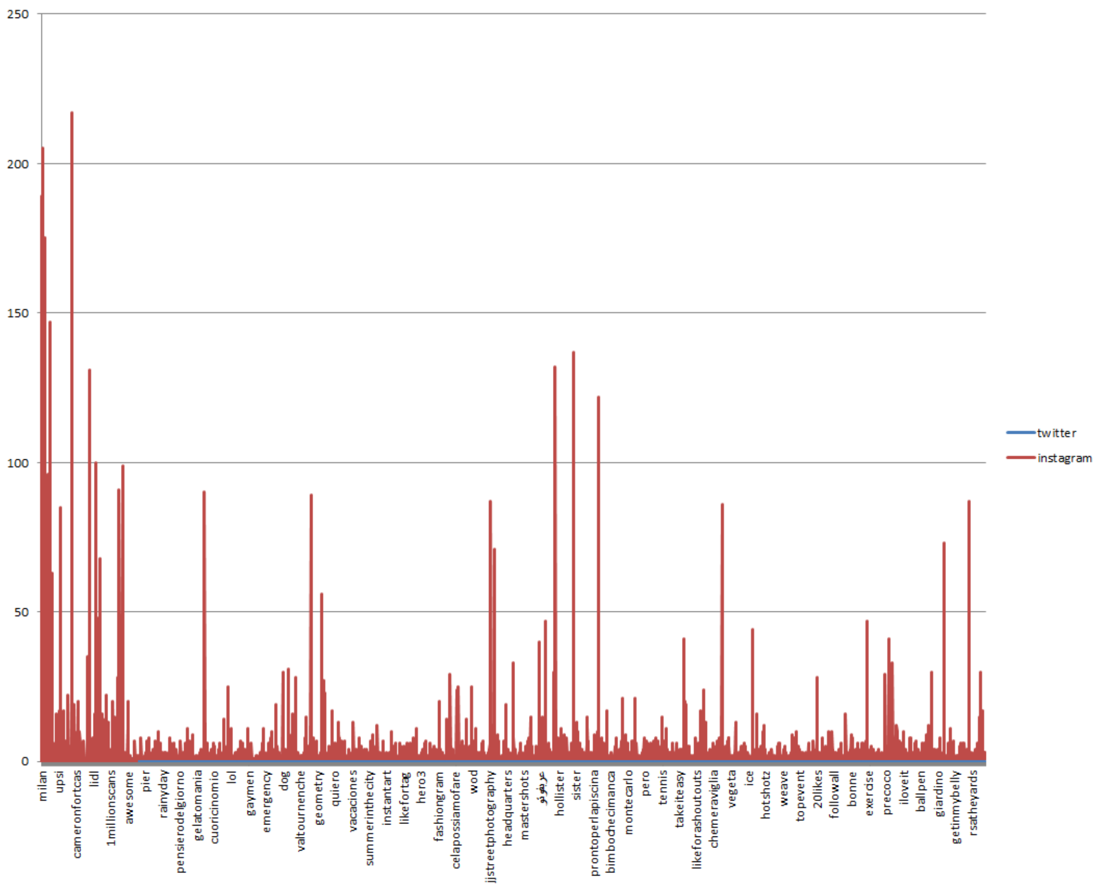


FIGURE 6.1: The graph for occurrences of bag of tags

6.1.3 Identifying Correlation Strength

The next step was to elicit the strength of the relation between the two platforms, according to these results. It has been decided to check for the Pearson correlation coefficient which is a parameter to evaluate the linear correlation between two different equal sized datasets.The Java Apache Common Statistics library was used.

Results

Pearson Correlation Coefficient : 0.2950

Discussion of Results

The result was not very satisfactory having the value around 0.3, which demonstrates a weak correlation. It is known that the values of correlation coefficient having the absolute value between 0 and 0.3 is a weak, 0.3 to 0.5 a sign of medium correlation and values being greater than 0.5 is a representation of a strong relation.

Even before calculating the hashtags used per posts for Instagram and Twitter, it is possible to identify that there is a remarkable difference of this number between the platforms. Instagram entries are containing more occurrences of hashtags per entry. This fact is considered as the reason of the low correlation coefficient. Pearson analysis is a parameter of linear correlation, but taking into consideration of this unbalance of hashtags it might be inefficient to search for a linear relation. For this reason it has been decided to use Spearman's Rank Correlation analysis which is able to detect non-linear relations.

6.2 Analysis 2

6.2.1 Deciding on Evaluation Parameter

After the analysis of coefficient of correlation, It is considered that it may be more efficient to use Spearman's Rank Correlation Coefficient which is a non-parametric version of Pearson. This evaluation can be used when there is an assumption of a monotonic relationship and gives an idea about the correlation between the ranks instead of the values.

Results

With this evaluation technique, the value 0.53, for correlation coefficient was acquired.

Discussion of Results

It can be stated that the value 0.53 is a demonstration of a roughly strong relationship. And as it can be observed, almost 100% increase is achieved in this calculation with respect to Pearson. It is possible to state that there is non-linear, and linear components of the correlation; linear being weak and non-linear being strong.

6.2.2 Identifying the stopwords and the data cleaning process

As it has been stated before in the discussions, after the previous analysis some issues affecting the correlation coefficient and grouping has been identified. To overcome these issues it was important to clean the stop-words, and platform specific tags (which has been identified to give no information about the content of the post). For the cleaning process there were two options to consider:

- To exclude these tags from the analysis
- Use inverse document frequency weighting approach.

For the sake of simplicity, it is chosen to exclude the tags that are containing the stop-words and also platform specific key-words.

Here are the list of stop-words, platform specific words that have been used: “in”,“for”,“at”,“or”,“f4f”,“l4
all tags containing: “like”,“oftheday”, “follow”,“vsco”,“iger”,“insta”,“stagram”

There has been also the issue for the different grouping for the tags implying the same concept/topic because of the syntax error and language:

The solution that has been offered to overcome this issue was to identify these tags and map them to one unique tag before the grouping operation. Here are a few examples from this mapping:

“milano”: “milan”,

“parigi”: “paris”,

“londra”: “london”,

“roma”: “rome”,

“italia”: “italy”,

“italya”:“italy”,

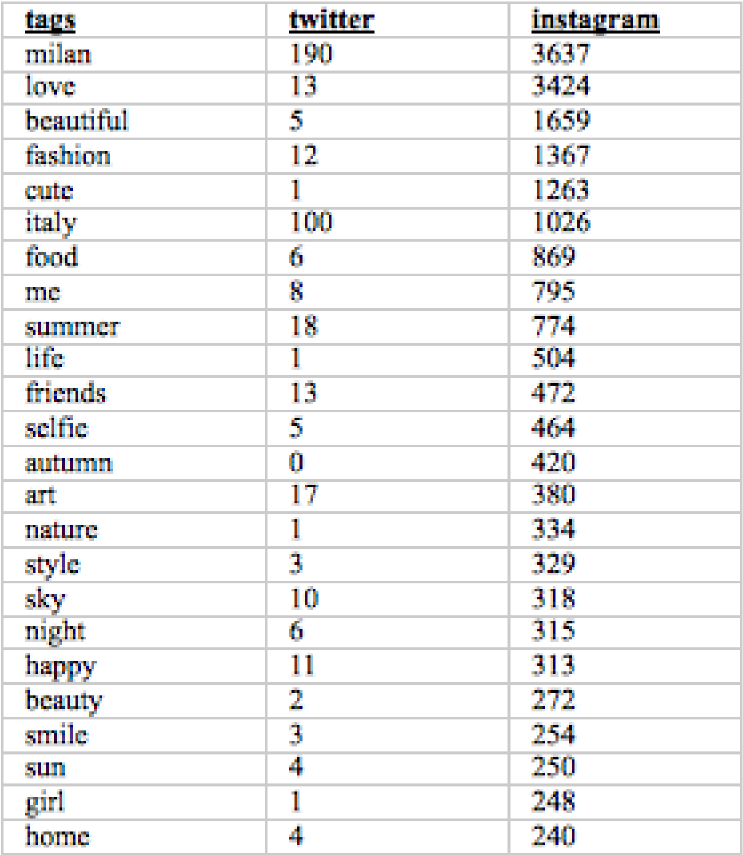
“francia”:“france”

Results

An example form top 100 tags for the week 11.11.2014 tags

Pearson : 0.34

Spearman : 0.61



tags	twitter	instagram
milan	190	3637
love	13	3424
beautiful	5	1659
fashion	12	1367
cute	1	1263
italy	100	1026
food	6	869
me	8	795
summer	18	774
life	1	504
friends	13	472
selfie	5	464
autumn	0	420
art	17	380
nature	1	334
style	3	329
sky	10	318
night	6	315
happy	11	313
beauty	2	272
smile	3	254
sun	4	250
girl	1	248
home	4	240

Table 6.2: Top 100 popular tags excluded PS tags

Discussion of Results

Comparing the correlation coefficients before and after the cleaning of platform specific tags and stop words one can recognize the increase in both Pearson and Spearman values. From this observation it is possible to infer that the correlation has both linear (weak) and non-linear (strong) components, and the quality increases as the noise is cleaned, as expected.

6.3 Analysis 3

Correlation coefficient for different granularities in the Time interval

After the analysis for one week from 11.11.2014 to 18.11.2014, for the next analysis, different granularities in the whole time interval has considered and the correlation between the topics is analyzed. Monthly, weekly and daily correlation analysis has been

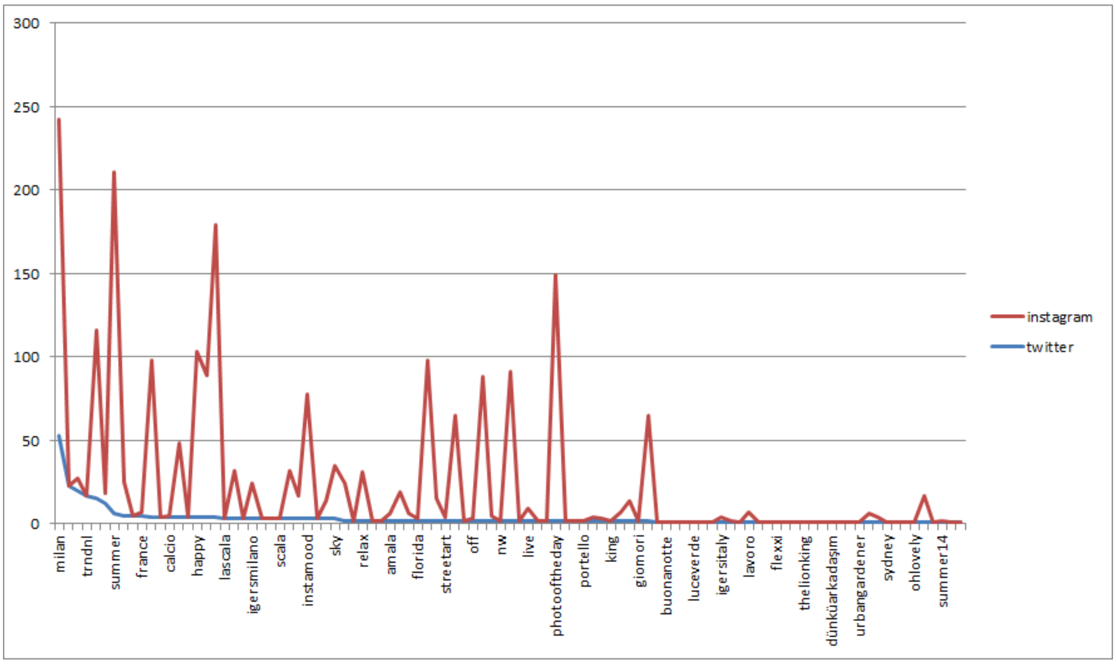


Figure 6.2: Shared tag percentages for top 100 between 2 platforms

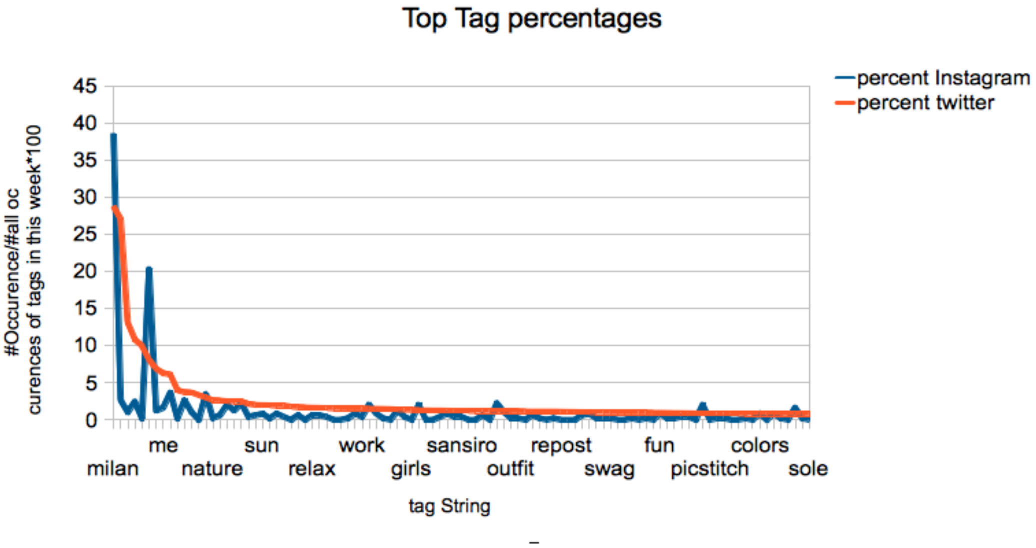


Figure 6.3: Top Tag Percentages for the week

conducted and percentage of shared tags are calculated between top 100 hashtags in each platform for each granularity.

Results

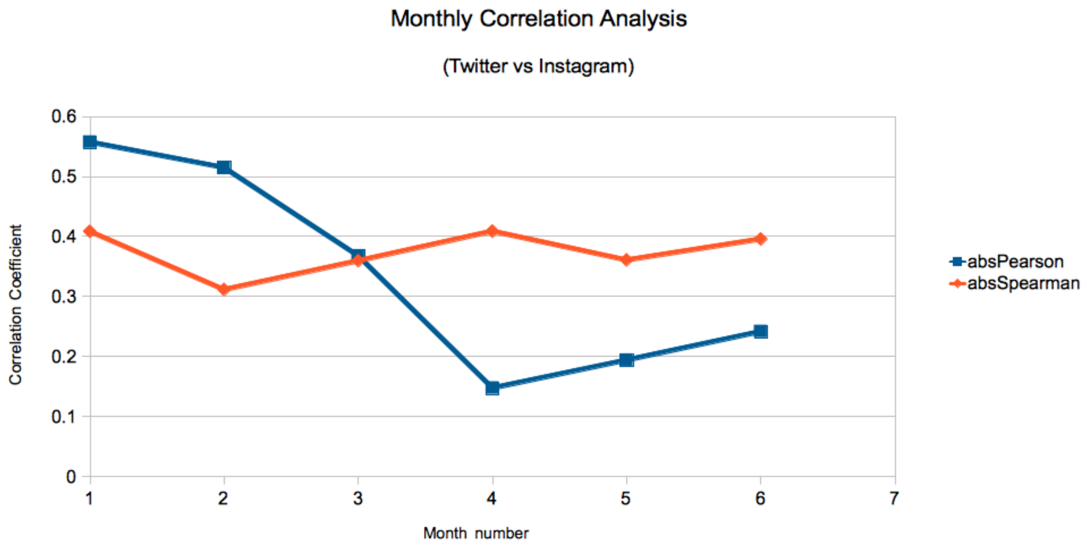


FIGURE 6.4: Monthly Correlation Analysis

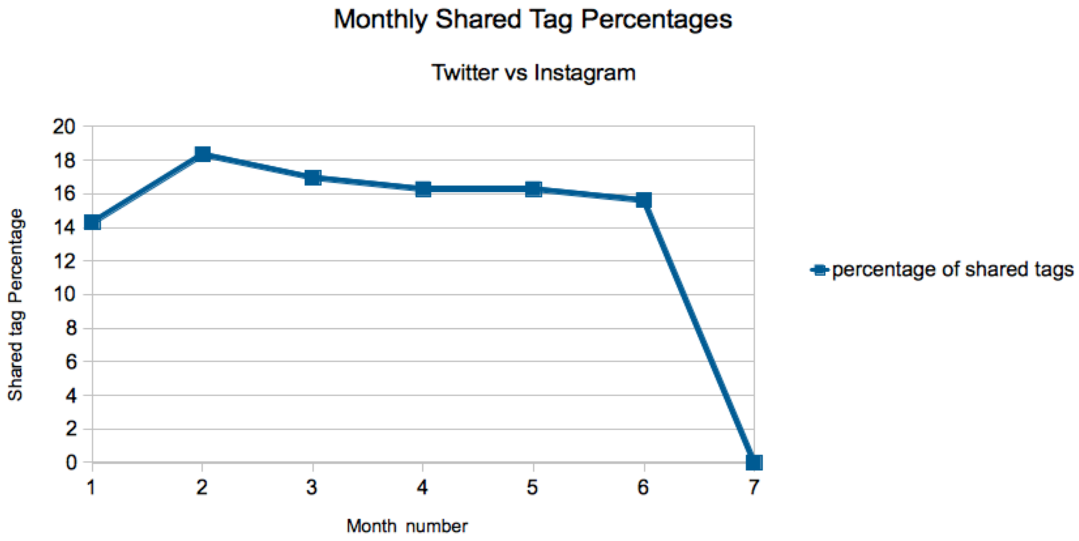


FIGURE 6.5: Monthly Top Shared Tag Percentages

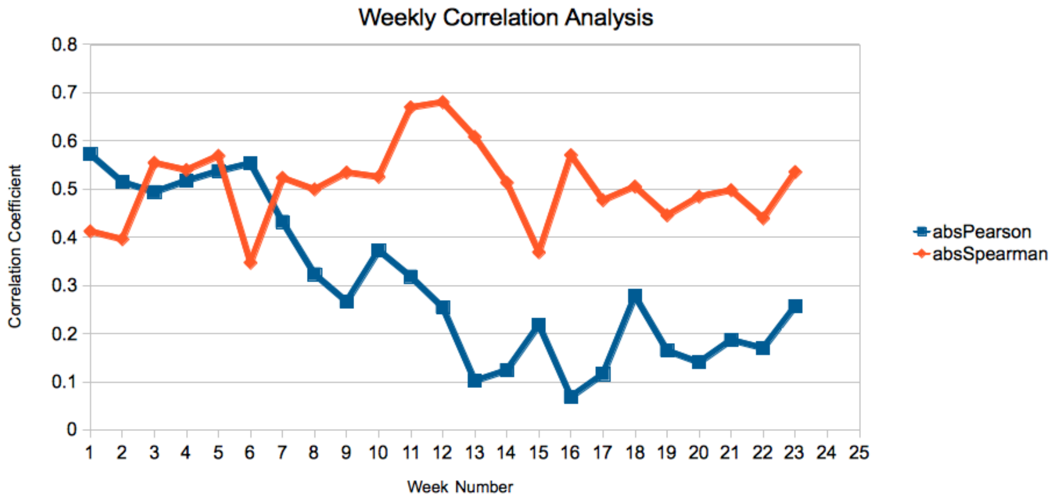


FIGURE 6.6: Weekly Correlation Analysis

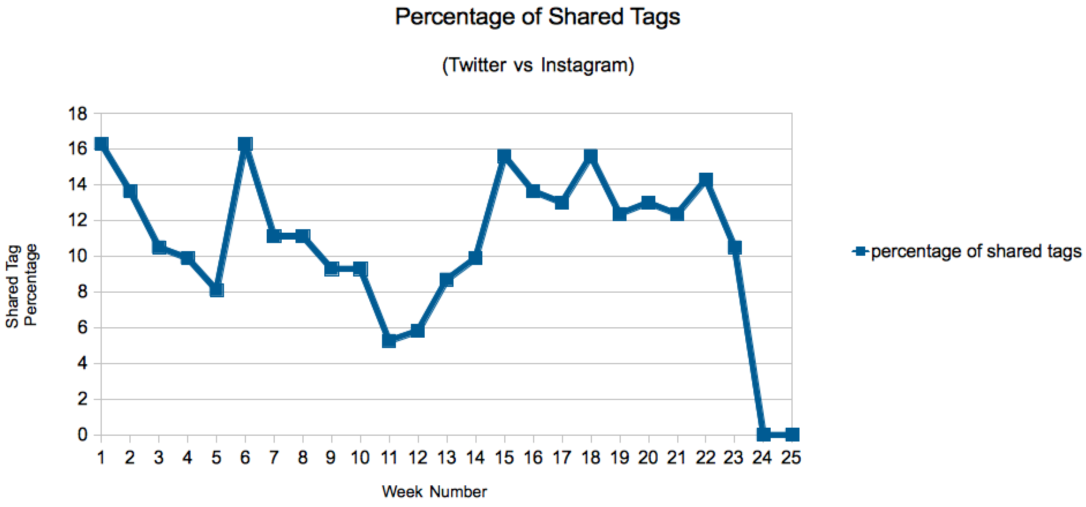


FIGURE 6.7: Weekly Top Shared Tag Percentages

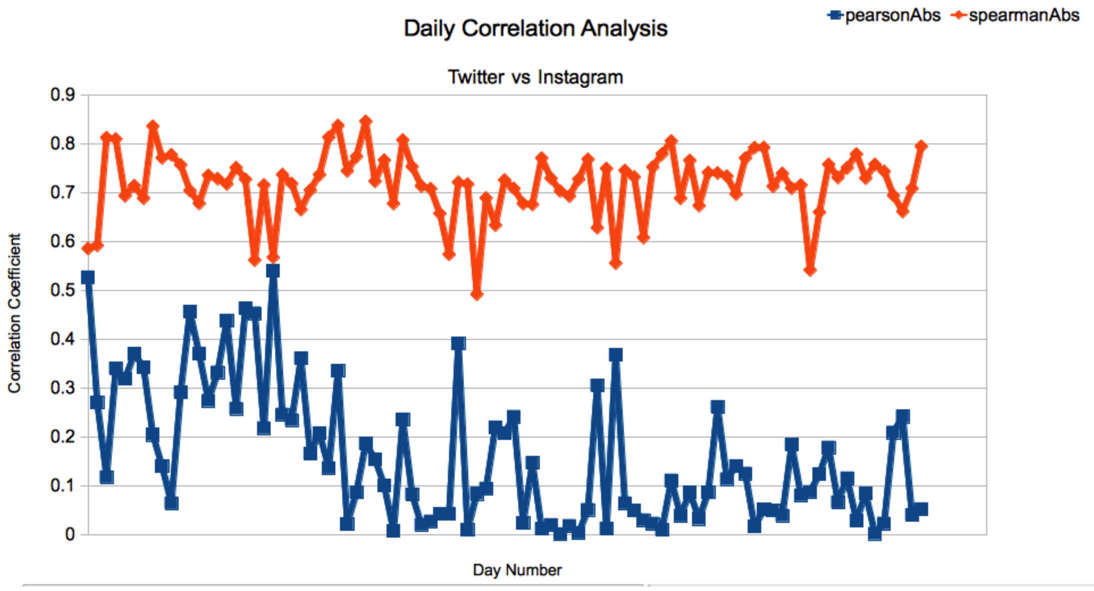


FIGURE 6.8: Daily Correlation Analysis

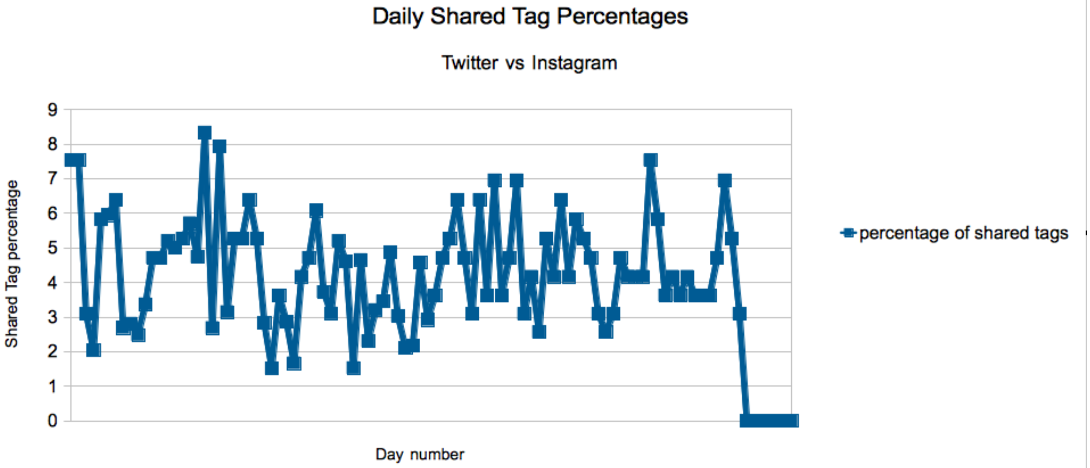


FIGURE 6.9: Daily Top Shared Tag Percentages

Discussion of Results

It has been observed that especially in a daily basis, the correlation between topics in Instagram and Twitter is pretty strong, the values are changing from 0.7 to 0.9. Having the correlation coefficient more than 0.6, the weekly relation of the top 100 topics between Instagram and Twitter is also pretty high. From this observation, it can be concluded that if a topic is popular in Twitter in a particular week or day it is very probable that this topic is also popular in Instagram during the same time. Furthermore, it is possible to observe that the shared tag percentages increase and decrease proportionally with the correlation coefficients, this shows consistency between correlation and shared tags percentage as expected.

6.4 Analysis 4

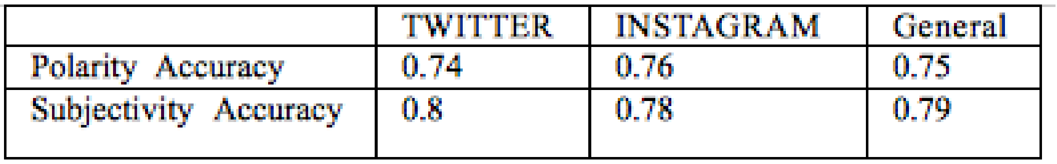
Sentiment Analysis in each Platform

After examination of the number of shared topics and tags in different platforms and observing the correlation coefficient values, in the next analysis it was interesting to search for a relationship between the sentiments for the entries between different platforms. For this goal, the new dataset prepared containing sentiments for each entity is used.

6.4.1 Accuracy analysis for the Sentiment Functionality

Before doing the sentiment analysis and start using the results it was very necessary to evaluate the accuracy of the web service's sentiment analysis functionality. There were few options to do this: to use a sentiment corpus, or to take a random sample set from the dataset, manually label them, and make an evaluation.For the sake of consistency with the datasets since they are not usual texts but social network posts, the last option was preferred. 50 English posts from Twitter and 50 English posts from Instagram has selected randomly and labeled as positive neutral or negative as well as subjective or objective. Then the evaluation of accuracy has been conducted.

Results



	TWITTER	INSTAGRAM	General
Polarity Accuracy	0.74	0.76	0.75
Subjectivity Accuracy	0.8	0.78	0.79

Table 6.3: Evaluation of Accuracy for Sentiment Analysis

Discussion of Results

It has been observed that having the values more than 0.75 the functionality is accurate enough to be used for following steps.

6.4.2 General Analysis for Polarity and Subjectivity of Sentiments

It is important to state that for this analysis as the sentiment function is available for only the English language, the dataset used for the following parts is covering only the entries that are identified to be in English.

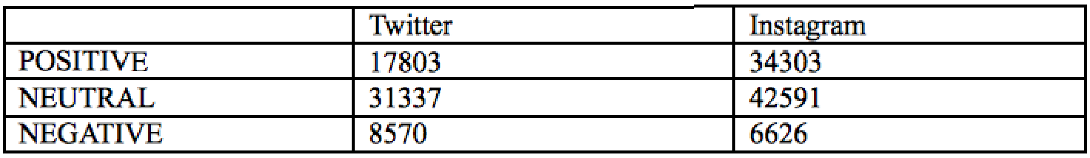
The new datasets without considering any granularity or subjects have been analyzed according to subjectivity objectivity rates as well as positivity neutrality and negativity.

Covered:

- 32797 tweets
- 81113 instagram posts.

  Summing up to 113.910 entries to be analyzed.

Results



	Twitter	Instagram
POSITIVE	17803	34303
NEUTRAL	31337	42591
NEGATIVE	8570	6626

Table 6.4: Evaluation for General Polarity of Sentiments

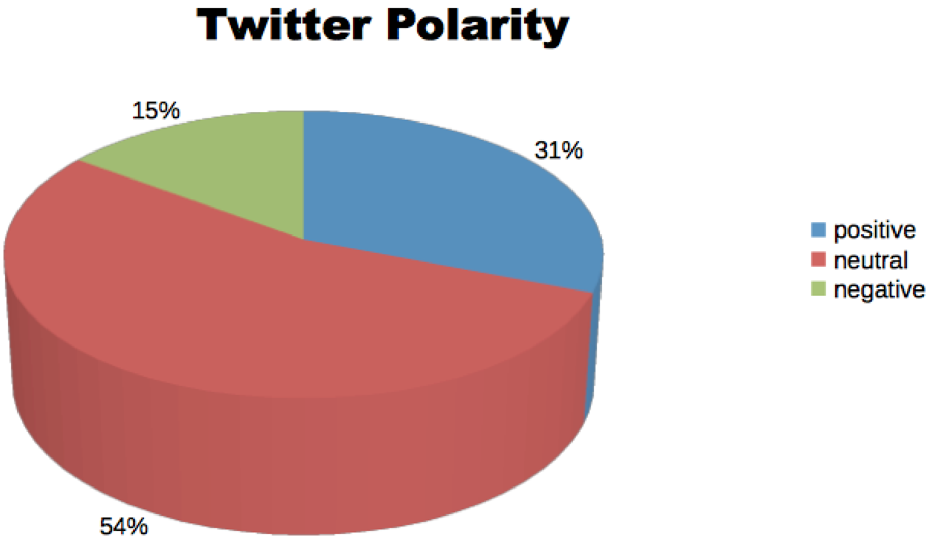


Figure 6.10: General Polarity of Sentiments for Twitter

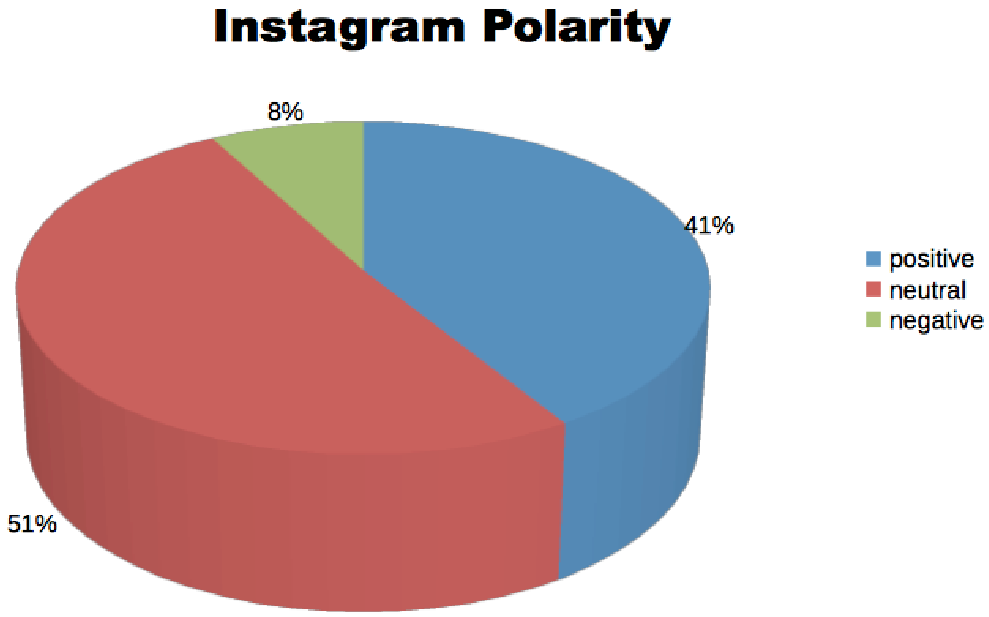
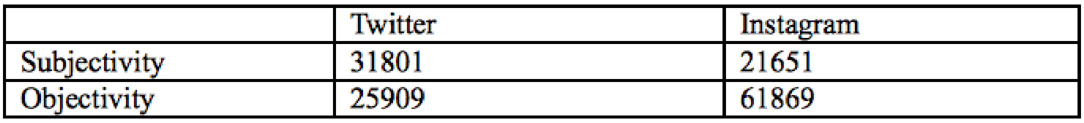


FIGURE 6.11: General Polarity of Sentiments for Instagram



	Twitter	Instagram
Subjectivity	31801	21651
Objectivity	25909	61869

TABLE 6.5: Evaluation for General Subjectivity of Sentiments

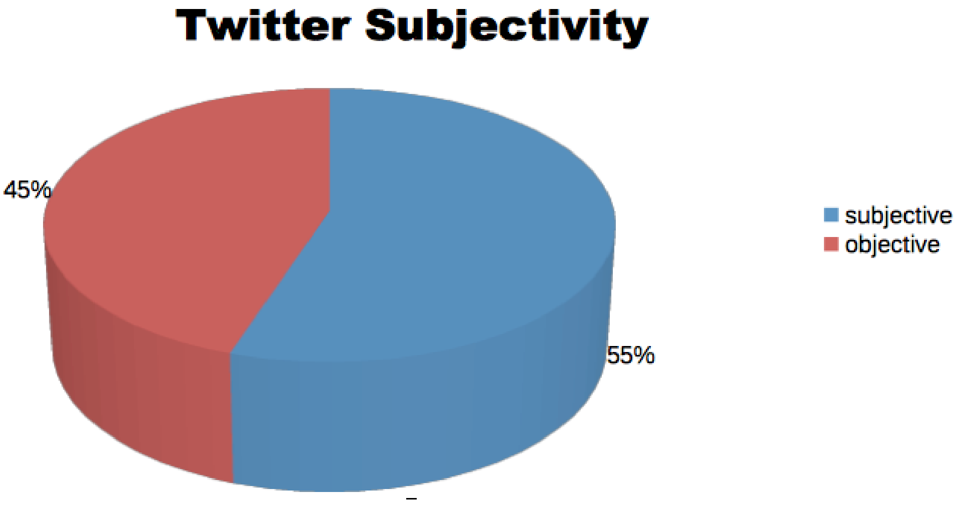


FIGURE 6.12: General Subjectivity of Sentiments for Twitter

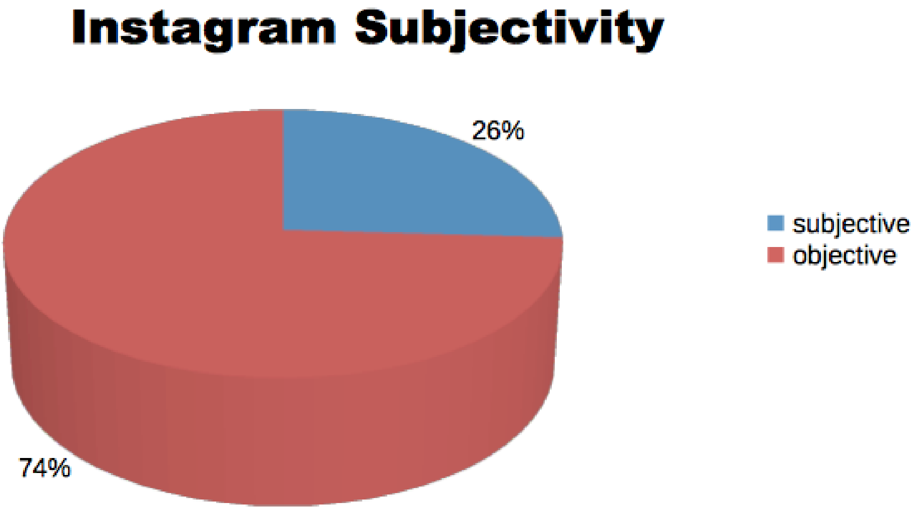


FIGURE 6.13: General Subjectivity of Sentiments for Instagram

Discussion of Results

According to the results about the general polarity of the sentiments in Instagram and Twitter it is possible to see that the neutrality percentage is almost the same in both platforms. On the other side about negativity, one can observe that the percentage is higher in general in Twitter 6.10 by being almost the twice as much as the percentage of negativity in Instagram 6.11. From this observation, it can be deduced that, although in both platforms the main tendency of the posts is positive and neutral, it is more probable to encounter a negative post in Twitter compared to Instagram.

In addition it can be observed that there is a remarkable difference in terms of subjectivity in Twitter 6.12 and Instagram 6.13. Twitter tend to show more subjectivity more than twice of the percentage in Instagram. Taking into account both of these observations about the negativity and subjectivity one can infer that in Instagram people share what they think positively about (as we know with the photos) and tend to be positive about them so there are a few negative entries, and in Twitter people tend to make more personal comments, share subjective ideas, which may also contain negativity.

6.5 Analysis 5

Weekly sentiment analysis of top 100 tags in Instagram and Twitter

Following the analysis of the correlation between different platforms for top 100 tags, it has been considered to be interesting to get into details and examine the sentiments of these posts containing top 100 hashtags and see if there is a coherency between Instagram and twitter, by means of sentiments.

For this part weekly granularity has been chosen to examine sentiments of top 100 subjects.

Results

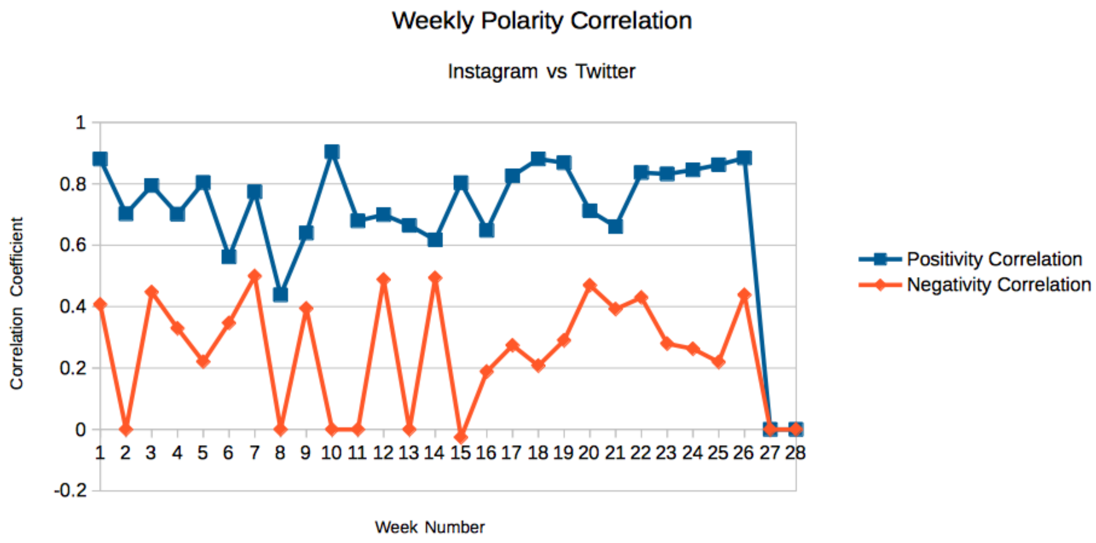


FIGURE 6.14: Weekly Correlation of Polarity

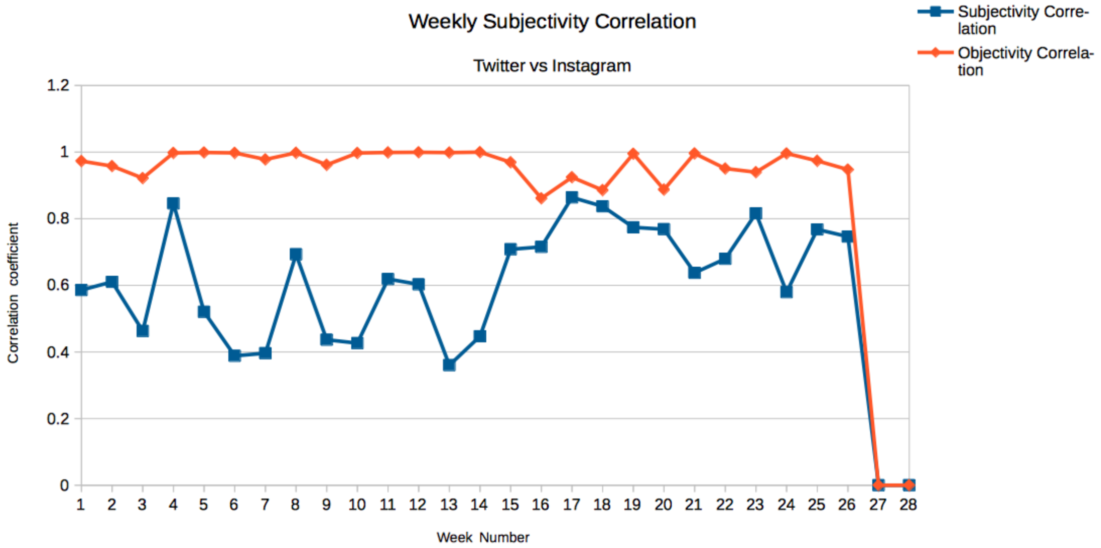


FIGURE 6.15: Weekly Correlation of Subjectivity

Discussion of Results

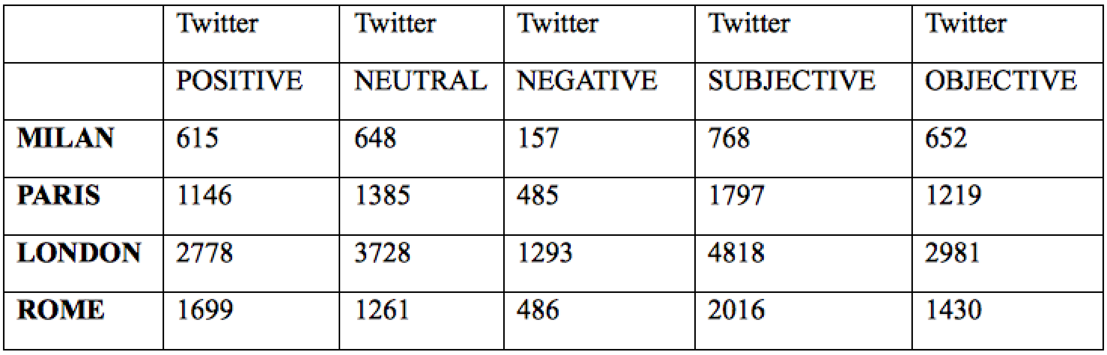
As the results have been examined and correlation coefficients have been evaluated one can state that the there is a strong correlation, especially between positivity and the objectivity of the sentiments between Instagram and Twitter. From this fact it is possible to deduce that if there is a topic that has been mentioned in one week in Twitter positively, it is very probable that in the same week, the instagram posts mentioning this topic will have a positive sentiment. About negativity and subjectivity it is consistent with the previous discussion 6.4.2 that the correlation coefficient values are lower, since generally in Twitter there is a tendency of having more subjective and negative posts than in Instagram.

6.6 Analysis 6

Analyzing the general polarity and subjectivity of sentiments for the posts mentioning 4 big cities, in Instagram and Twitter

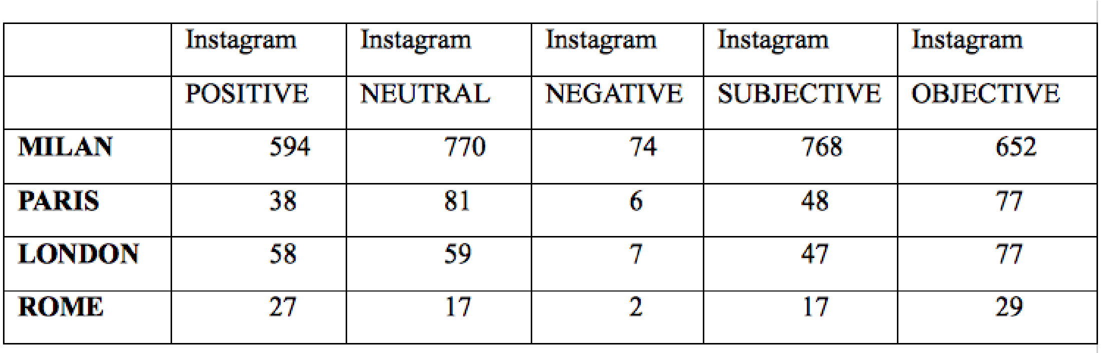
To confirm the inference that has been made in analysis 6.5 which mainly states that if some topic is mentioned positively in one platform there is a high probability that it will be mentioned also positively in other platform, as well as checking the induction about the tendency of Twitter in general containing subjective and Instagram objective posts which has been deduced from analysis 6.4.2, it has been decided to focus on the 4 big cities and analyze the polarity and subjectivity of the sentiments of the posts mentioning these cities in each platform.

Results



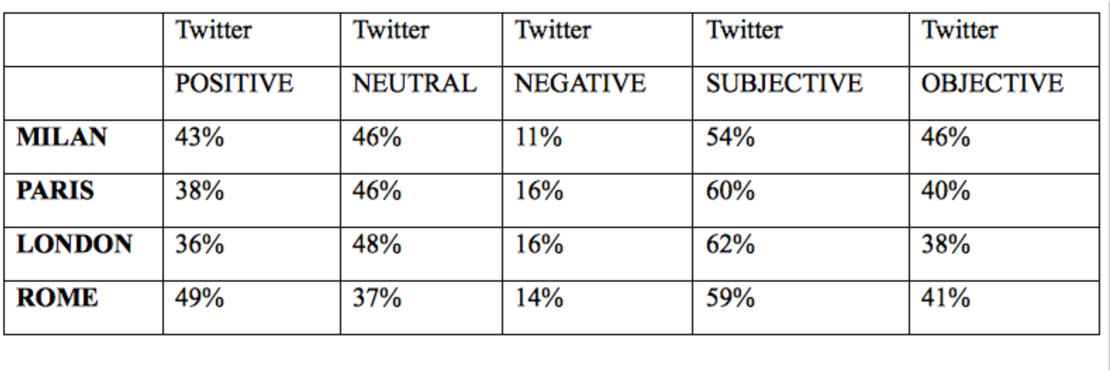
	Twitter	Twitter	Twitter	Twitter	Twitter
	POSITIVE	NEUTRAL	NEGATIVE	SUBJECTIVE	OBJECTIVE
MILAN	615	648	157	768	652
PARIS	1146	1385	485	1797	1219
LONDON	2778	3728	1293	4818	2981
ROME	1699	1261	486	2016	1430

Table 6.6: Sentiment results of Twitter for 4 big cities



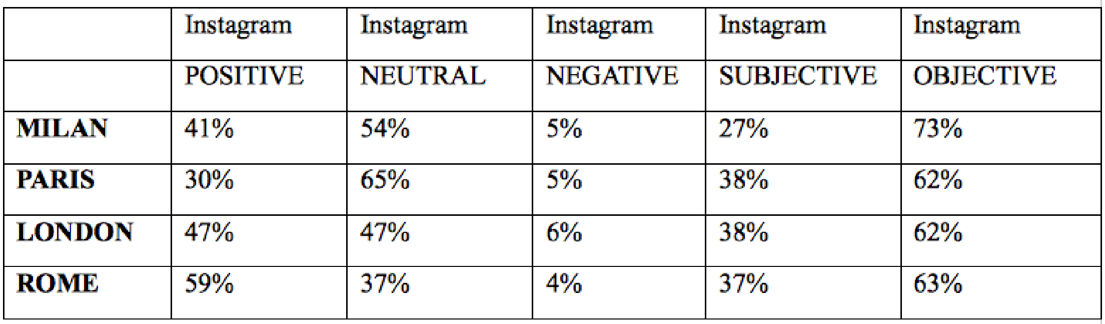
	Instagram	Instagram	Instagram	Instagram	Instagram
	POSITIVE	NEUTRAL	NEGATIVE	SUBJECTIVE	OBJECTIVE
MILAN	594	770	74	768	652
PARIS	38	81	6	48	77
LONDON	58	59	7	47	77
ROME	27	17	2	17	29

Table 6.7: Sentiment results of Instagram for 4 big cities



	Twitter	Twitter	Twitter	Twitter	Twitter
	POSITIVE	NEUTRAL	NEGATIVE	SUBJECTIVE	OBJECTIVE
MILAN	43%	46%	11%	54%	46%
PARIS	38%	46%	16%	60%	40%
LONDON	36%	48%	16%	62%	38%
ROME	49%	37%	14%	59%	41%

Table 6.8: Sentiment rates of Twitter for 4 big cities



	Instagram	Instagram	Instagram	Instagram	Instagram
	POSITIVE	NEUTRAL	NEGATIVE	SUBJECTIVE	OBJECTIVE
MILAN	41%	54%	5%	27%	73%
PARIS	30%	65%	5%	38%	62%
LONDON	47%	47%	6%	38%	62%
ROME	59%	37%	4%	37%	63%

Table 6.9: Sentiment rates of Instagram for 4 big cities

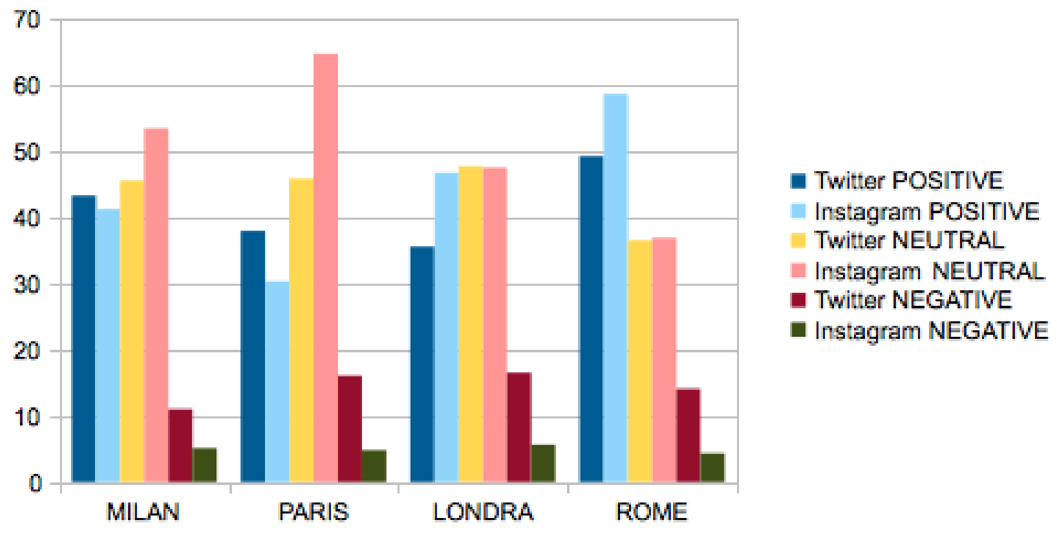


Figure 6.16: Polarity percentages of city mentions

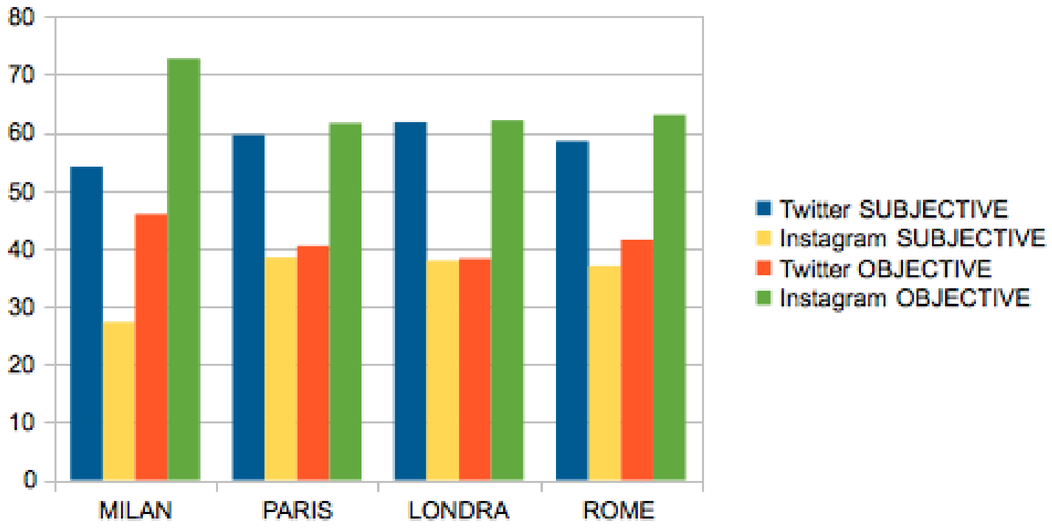


Figure 6.17: Subjectivity percentages of city mentions

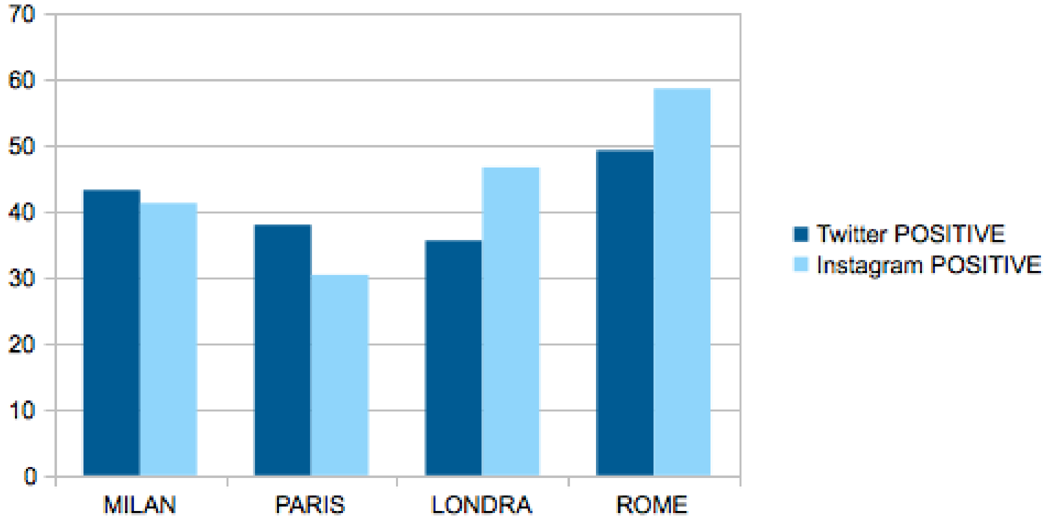


Figure 6.18: Positivity percentages of city mentions

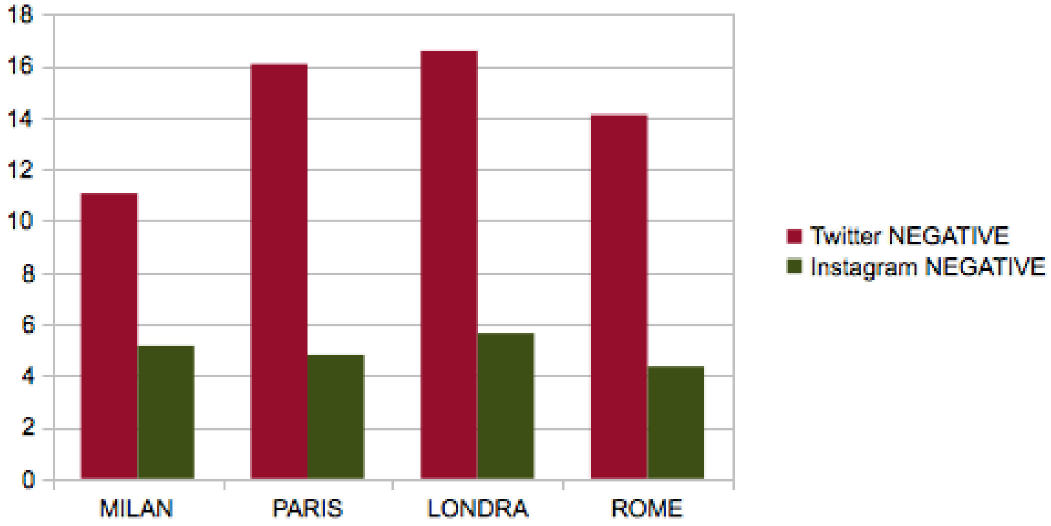


Figure 6.19: Negativity percentages of city mentions

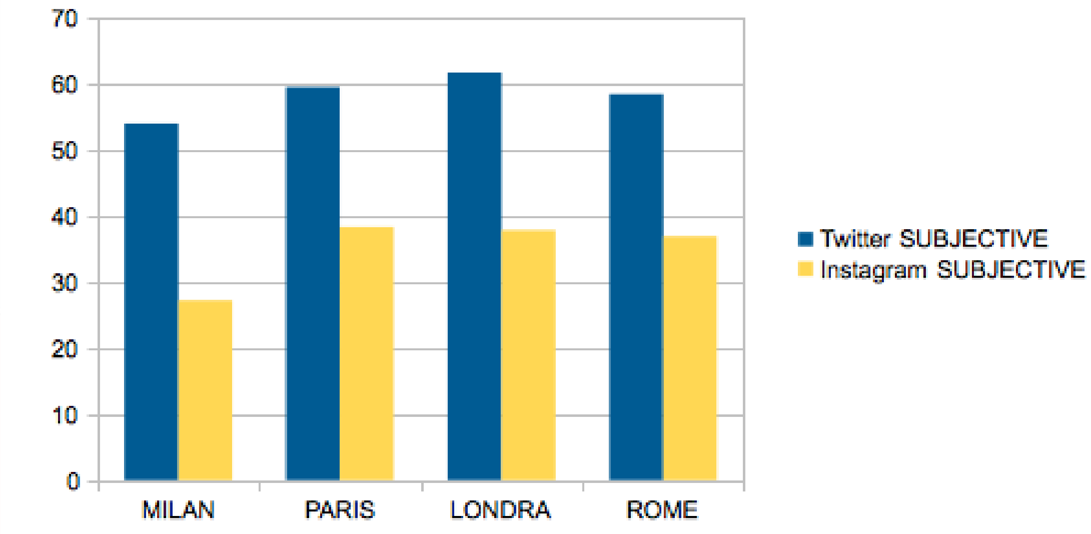


Figure 6.20: Subjectivity percentages of city mentions

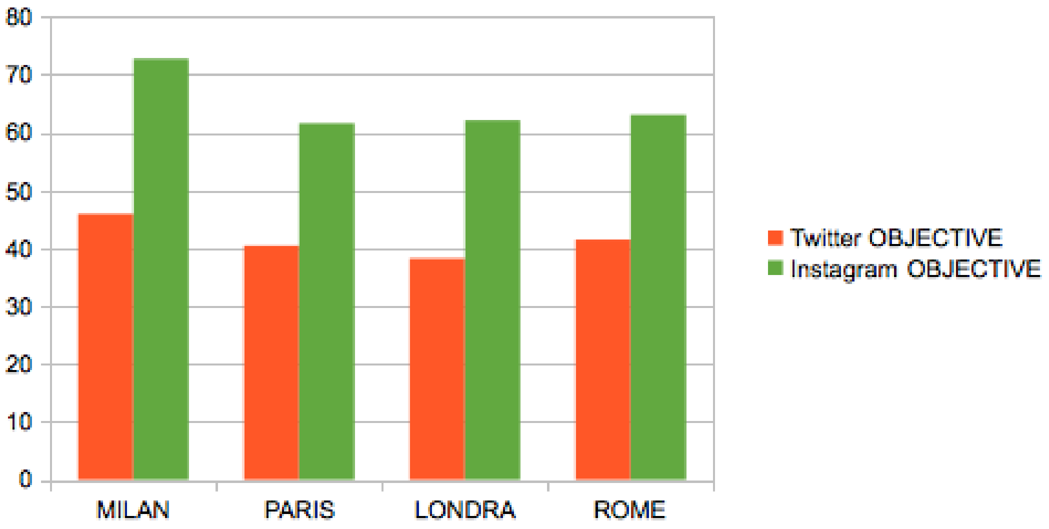


Figure 6.21: Objectivity percentages of city mentions

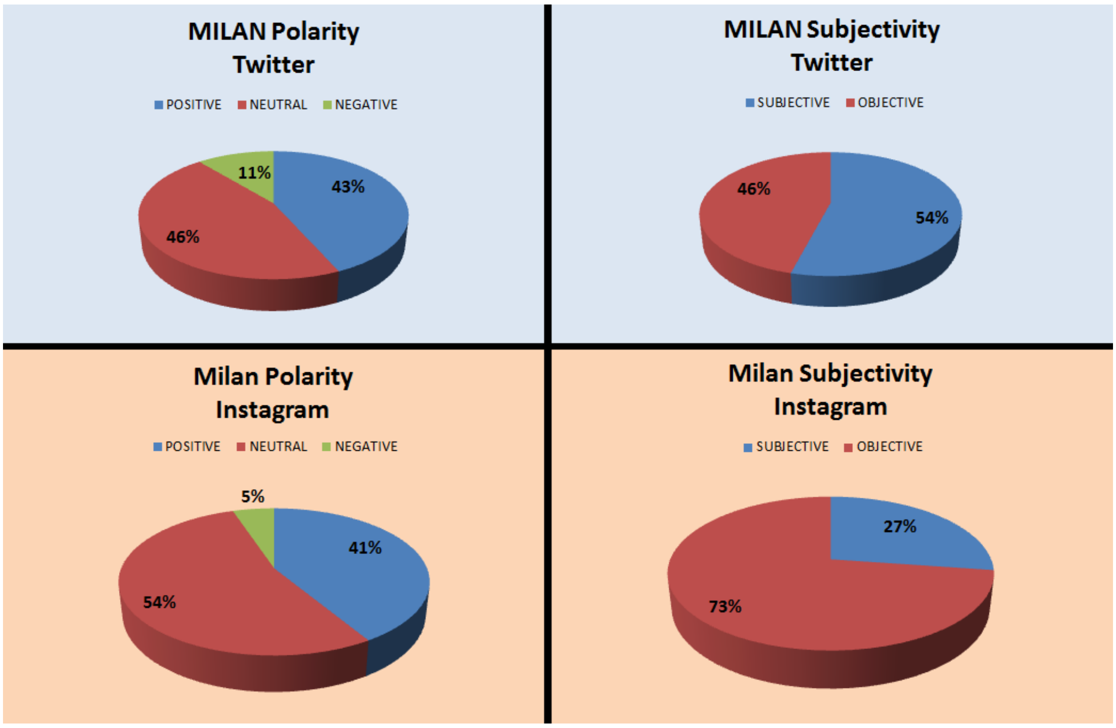


Figure 6.22: Polarity subjectivity percentages for Milan mentions

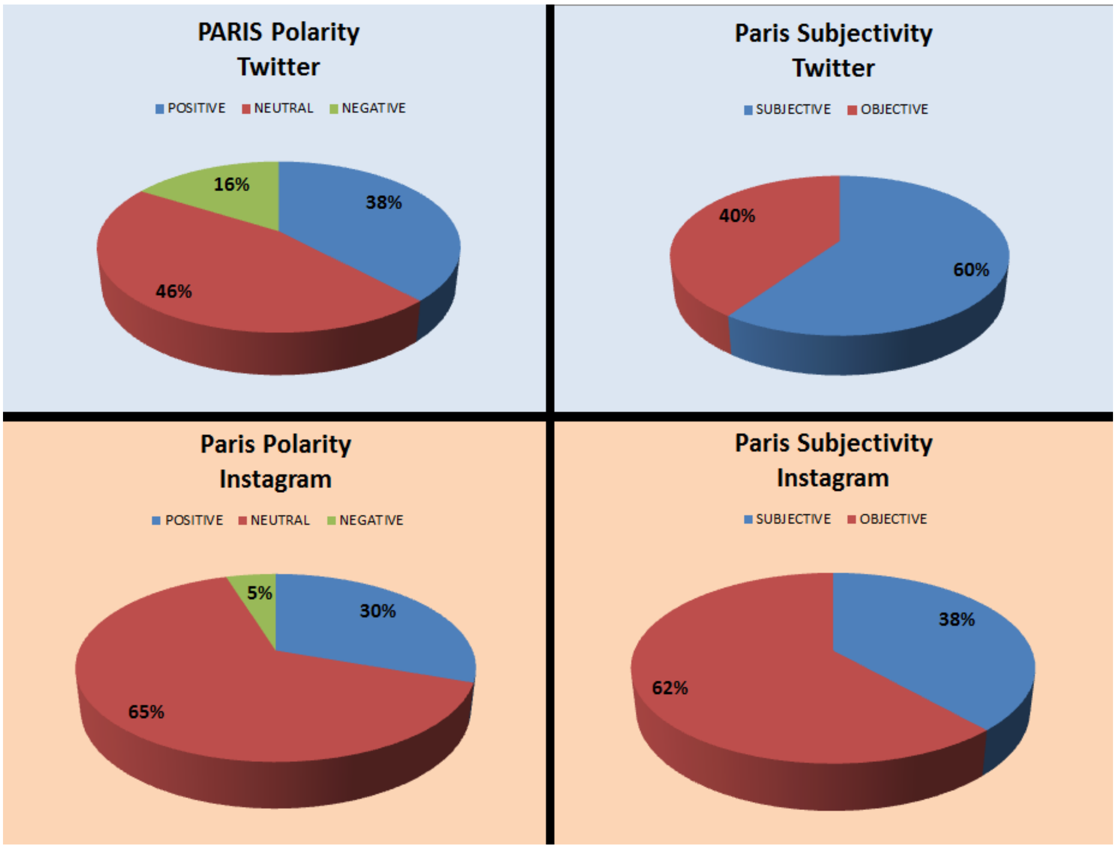


Figure 6.23: Polarity subjectivity percentages for Paris mentions

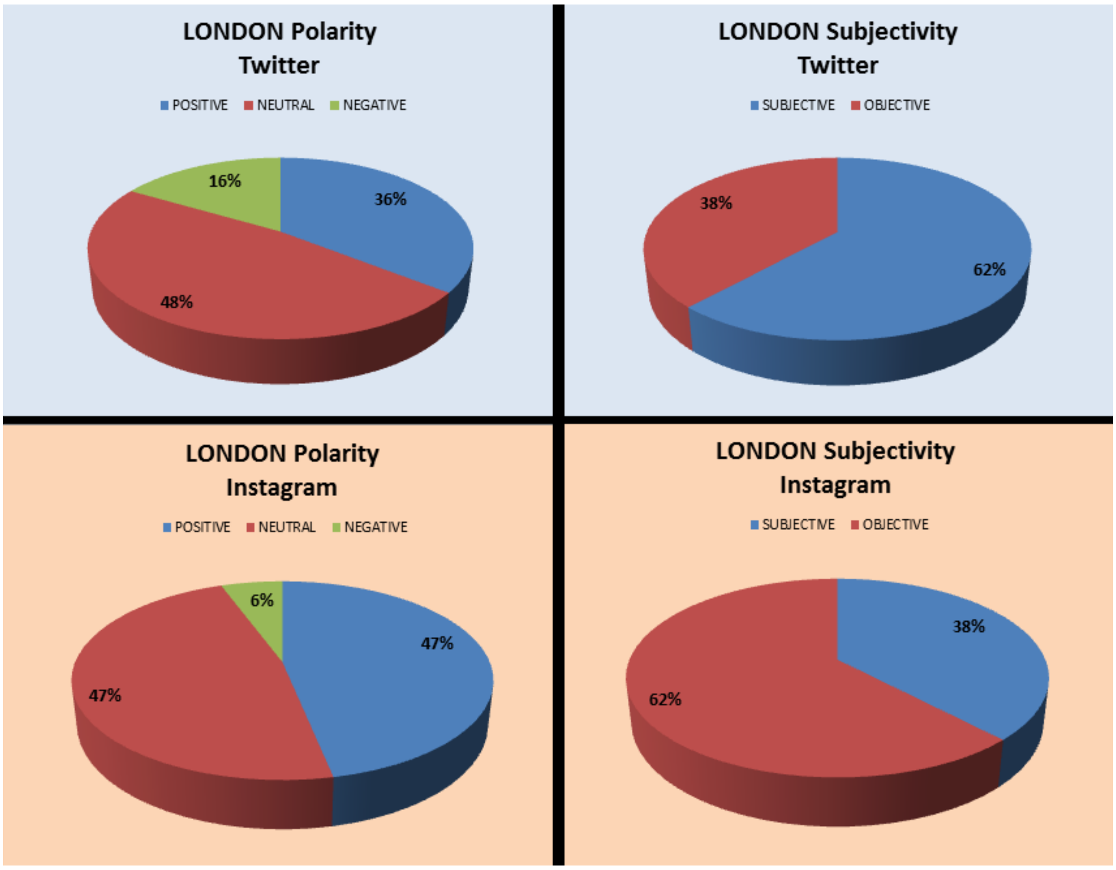


Figure 6.24: Polarity subjectivity percentages for London mentions

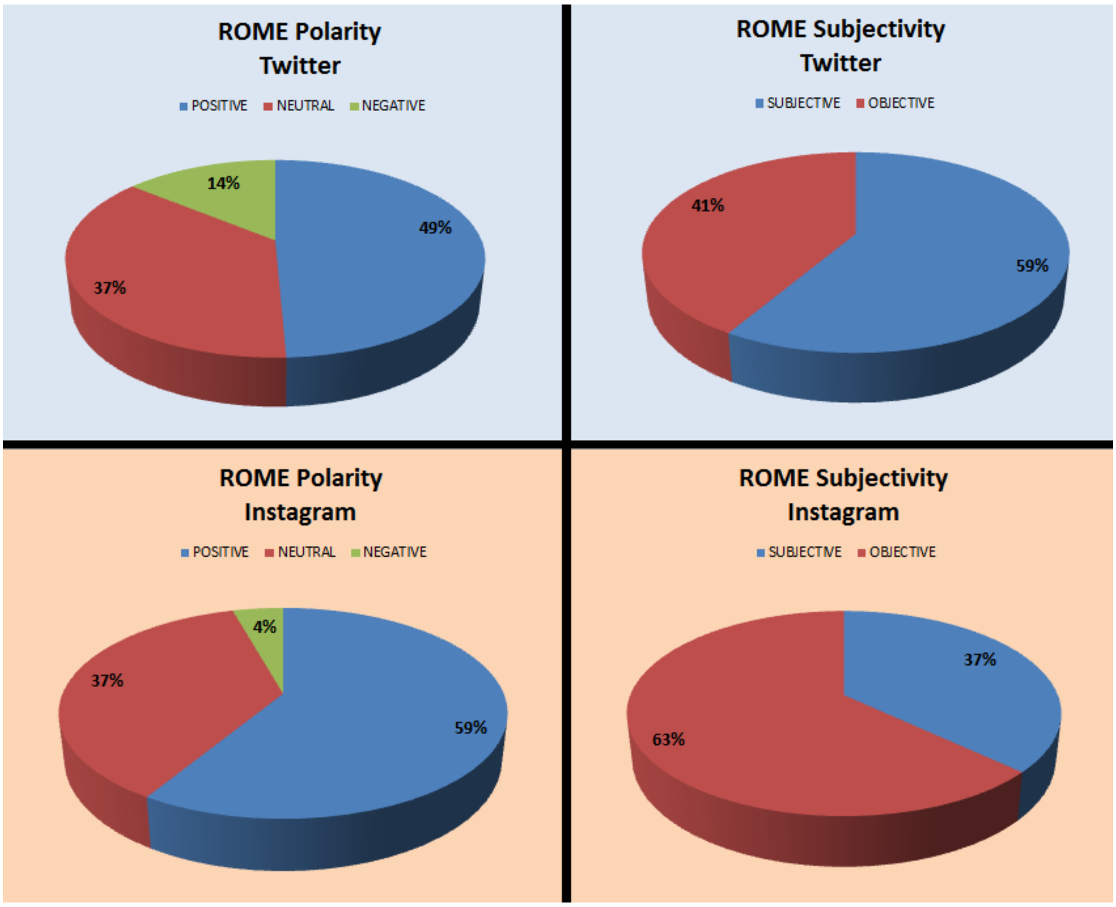


FIGURE 6.25: Polarity subjectivity percentages for Rome mentions

Discussion of Results

Concerning the subjectivity aspect it can be seen from Figure 6.17 that, the results in this analysis confirms the induction that has been made about the subjectivity in 6.4.2. Concerning all of the 4 cities, in Instagram, the objectivity has a greater percentage than 60%. In addition holding the same idea, in Twitter, subjectivity has a higher percentage than 50% for all the 4 cities.

Concerning the polarity, again confirming the previous inferences 6.4.2, it can be seen that the posts tend to be neutral in both platforms for all 4 cities. On the other side, if the values of percentages are compared in terms of negativity and positivity, it is possible to see the difference between Instagram and Twitter. For all of the 4 cities the negativity percentage in Twitter is almost 3 times asmuch as the value of negativity percentage in Instagram. 6.19 So it can be told that the neutrality is around the same percentage comparing 2 platforms for one city 6.16, but the ratio of negativity/positivity is much higher in Twitter than in Instagram, and this fact counts for both 4 cities.

- Milan negativity/positivity ratio :
  - Twitter : $\frac{11}{43}$=0.2558
  - Instagram : $\frac{5}{41}$=0.122
- Paris negativity/positivity ratio:
  - Twitter : $\frac{16}{38}$=0.421
  - Instagram : $\frac{5}{30}$=0.166
- London negativity/positivity ratio:
  - Twitter : $\frac{16}{36}$=0.444
  - Instagram : $\frac{6}{47}$=0.127
- Rome negativity/positivity ratio:
  - Twitter : $\frac{14}{49}$=0.285
  - Instagram : $\frac{4}{59}$=0.067

Hence it can be stated that it is much more probable to encounter subjective and negative comments in Twitter than in Instagram, which confirms the previous inductions 6.4.2. For this reason if one would like to make a sentiment analysis for an entity as brand, person or so on, Twitter would give more realistic and subjective results compared to Instagram.

Chapter 7

Conclusions

This chapter is dedicated to give the summary of the entire work done, discussion of the results achieved, and the possible future work.

7.1 Summary

The main significant result of the work was that it offered an approach for showing the dynamics and relation of content between different platforms. Different parameters are identified for seeking for the relation, ranging from the most popular topics, sentiments to covering specific granularity of time intervals and mentioned entities. Considering the formats of the social network posts, an approach covering the phases for extraction, processing, and analysis of this data has been proposed. Different aspects were considered to be taken into account while analyzing the data:

- The sentiments. The social network posts needed a human evaluation for understanding sentiments. The automatic labeling concerning sentiments using web service [3] has performed.
- The place names mentioned. The names of entities mentioned in texts are not easy to be identified automatically. The entity extraction is performed by using a web service[3].
- Most popular hashtags. Many authors use hashtags to demonstrate the main content of their posts. In other words, hashtags could be used for understanding the contents of the posts. Grounding on this fact the most popular tags are used for examining the relation between 2 platforms from the topics aspect.

The system has been executed on real data taken from Instagram and Twitter. The results and strength of the relation are evaluated for each aspect and represented to give a more clear idea about the dynamics of the pattern.

This thesis confirms that there is a remarkable relation between the posting behavior between different platforms (Instagram and Twitter), and it encloses some characteristic behaviors special to these different 2 platforms which differs from each other.

7.2 Critical Discussion

The concentration of the approach was on the relation of sentiments and most popular topics between Instagram and Twitter. Examining the relationship between different platforms was chosen as a main topic since until now there aren't any studies, focusing on this aspect. The results confirm that most popular hashtags have a strong relationship in a time interval between the platforms Instagram and Twitter. In other words, they tend to have the same popular topics in the same intervals(6.3). In addition concerning to sentiments this work has demonstrated that the posts in Twitter tend to be more subjective, and higher in negativity than the posts in Instagram(6.4.2, 6.5).

At this point, it is useful to state that the parts for entity extraction and sentiment analysis are based on the assumption of these functionalities working in an accurate way. The manual accuracy check in detail about the sentiment analysis can be found in section 6.4.1.

7.3 Possible Future Work

The approach that is proposed in the thesis work can be a starting point for more extensive examinations concerning the relation between posting behaviors in different platforms.

The research can be expanded in many different aspects, as an example other platforms can be used for the analysis, Foursquare, Flickr as well as Facebook and Google+.

Different dimensions can be analyzed for relationships as most popular brands or people. Furthermore entities different than places, like organizations, people, can be examined concerning the sentiment. One other possible research can be searching the sentiments of the posts mentioning different places according to seasons to identify the most popular places for different seasons. The sentiment analysis can be added to previous work that creates relational maps according to 4 big cities [9].

Bibliography

[1] Network statistics. URL http://www.statista.com/statistics/219903/number-of-worldwide-social-network-users/.

[2] Language detection library website. URL https://code.google.com/p/language-detection/l.

[3] Aylien text api website. URL http://aylien.com/text-api.

[4] Joda time. URL http://www.joda.org/joda-time/.

[5] Jongo mongo-java-driver: with ease. URL http://www.joda.org/joda-time/.

[6] The commons math library - statistics - apache. URL http://commons.apache.org/math/userguide/stat.html.

[7] Gabriela Andreea Morar Cristina Ioana Muntean and Darie Moldovan. Exploring the meaning behind twitter hashtags through clustering. *Business Information Systems Workshops*, pages 231–242, 2012.

[8] Oren Etzioni Ritter, Alan and Sam Clark. Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining. *Proceedings of the 23rd international conference on World wide web*, pages 1104–1112, 12 Aug. 2012.

[9] Daniele Allevi and Luca Losa. Relevance, relations and topics of interest in cities based on content mining from social networks. Master's thesis, Politecnico di Milano, 2014.

[10] Jiwei Li and Claire Cardie. Timeline generation: Tracking individuals on twitter. *Proceedings of the 23rd international conference on World wide web*, pages 643–652, 2014.

[11] Alberto Chioda. Monitoring urban mobility with social networks. Master's thesis, Politecnico di Milano, 2013/2014.

[12] Alexa. alexa top 500 sites on the web. URL http://www.alexa.com/topsites.

[13] Ev williams on twitter's early years. URL http://www.inc.com/issie-lapowsky/ev-williams-twitter-early-years.html?cid=em01011week40day04b.

[14] United states securities exchange commission. URL http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm.

[15] 300 million: Sharing real moments. URL http://blog.instagram.com/post/104847837897/141210-300million.

[16] Kevin Dela Rosa. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 2011.

[17] Social media update 2014. URL http://www.pewinternet.org/2015/01/09/social-media-update-2014.

[18] Frequency of social media use. URL http://www.pewinternet.org/2015/01/09/frequency-of-social-media-use-2/.

[19] Lior Rokach Oded Maimon. *Knowledge Discovery and Data Mining*. The name of the publisher, 2010.

[20] Cornelius T. Leondes. *Database and Data Communication Network Systems*. Academic Press, Jul 9, 2002.

[21] Micheline Kamber Jian Pei Jiawei Han. *Data Mining: Concepts and Techniques*. 2011.

[22] Data integration and data transformation (in kdd). URL http://www.slideshare.net/farshadbadie/data-integration-and-data-transformation.

[23] Mark et al. Hall. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, pages 10–18, 2009.

[24] Mark et al. Hall. ext mining and natural language processing: introduction for the special issue. *SIGKDD Explor. Newsl. 7.1*, pages 1–2, 2005.

[25] Phuoc Tran-Gia Hoßfeld, Tobias and Maja Vukovic. Crowdsourcing: From theory to practice and long-term perspectives. 2014.

[26] Marco Brambilla Bozzon, Alessandro and Stefano Ceri. Answering search queries with crowdsearcher. *Proceedings of the 21st international conference on World Wide Web*, pages 1009–1018, 16 Apr. 2012.

[27] Mausam A. Ritter, S. Clark and O. Etzioni. Named entity recognition in tweets: An experimental study. *EMNLP*, 2011.

[28] Digital, social & mobile worldwide in 2015 - we are social. URL http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/i.

[29] B. Boehm. *A spiral model of software development and enhancement*, 21(10):61–72, 1988.

[30] Atlanmod json discoverer. URL http://atlanmod.github.io/json-discoverer/#/.