A framework for aligning social media and web analytics data to support search engine marketing

Ongan, M.

Award date:
2011

A Framework for Aligning Social Media and Web Analytics Data to Support Search Engine Marketing

Murat Ongan

July 2011

TU/e Technische Universiteit Eindhoven University of Technology

adversitement

A Framework for Aligning Social Media and Web Analytics Data to Support Search Engine Marketing

Author:
Murat Ongan

Supervisors:
Dr. Mykola Pechenizkiy
Ir. Guido Budziak

Eindhoven University of Technology
Department of Mathematics and Computer Science

A thesis submitted in fulfillment of the requirements for the degree of
Master of Science in **Computer Science & Engineering**

Assessment Committee:
Dr. Mykola Pechenizkiy (Database and Hypermedia Group, TU/e)
Ir. Guido Budziak (Adversitement B.V.)
Prof. Dr. Paul De Bra (Database and Hypermedia Group, TU/e)
Dr. Natalia Sidorova (Architecture of Information Systems Group, TU/e)

July 2011, Eindhoven

Abstract

Online business market has been growing very fast since the 90's. Consequently, companies have been paying more attention to areas such as web analytics, social media and online marketing as a part of their online activities. Despite the individual developments in these areas, the gaps between them are expanding.

This project intends to close the distance between social media with web analytics and search engine marketing. As a part of this effort, a generic social media measurement framework is suggested. This framework is designed to allow various types of data analysis by combining the data from social media and web analytics.

A reference implementation of the framework has been developed. The collected data can be explored by users with the web interface. The framework has also been integrated with two major web analytics solutions in the market. Namely, Google Analytics and Adobe SiteCatalyst.

The possibilities for making use of social media data to create and improve search engine marketing campaigns have been explored and two methods have been identified. These techniques enable marketers to use this framework as a marketing decision support system by introducing new types of analyses.

First, a methodology is suggested to create and manage search engine marketing keywords. Social media messages retrieved by the framework are used as a corpus to identify candidate terms and to build an ontology of concepts which can be used to generate keywords semi-automatically. The ontology learning part is built upon latent relational hypothesis. A pair-pattern matrix is built to apply classification algorithms to identify the relationships between the terms in an ontology.

Second, a hypothesis is introduced which suggests that, public interest on different products may change depending on the geographical region. According to this hypothesis, social media posts are used to detect regional interests on different products. It has been shown that, the results obtained by this method are positively correlated with regional product sales data. This method can be used to focus product advertisements on more interested regions to increase the gain of an advertisement campaign.

Acknowledgements

This thesis would not have been possible without Dr. Mykola Pechenizkiy, my thesis supervisor, who supported me during all phases of this project with his encouragement, supervision and feedback. He was also a key part of my entire master education with his courses, which constitute the basis of this work.

I am equally grateful to my company, Adversitement, which enabled me to do this project by hiring me as an intern, introducing me to their business area, creating a cozy work environment and providing me with all the necessary resources. I owe my gratitude to their employees who were always supportive and helpful when I needed. I would like to make a special reference to Guido Budziak, my advisor at Adversitement, for his invaluable guidance, inspiration and contribution in every step of this work from specifying the problem and scope of the project to writing this very document.

Besides my advisors, I would like to thank the rest of my thesis committee: Paul De Bra and Natalia Sidorova for their valuable time, comments, questions and assessment.

Regarding the web analytics data, I would like to thank Vodafone Netherlands for sharing their confidential information with me and for their support in academic research.

Lastly, I offer my thanks and regards to all of those, who made this work possible by supporting me in any respect. Special thanks go to my professors in TU/e and METU who taught me the principles of computer science; my roommates and close friends who never left me alone; and my dear family who raised, supported and loved me.

Murat Ongan

Contents

List of Tables

List of Figures

Acronyms

API application programming interface.

CPC cost per click.

CTR click-through rate.

HTTP hyper-text transfer protocol.

IR information retrieval.

LRA latent relational analysis.

LSA latent semantic analysis.

NLP natural language processing.

POS part-of-speech.

PR public relations.

ROI return on investment.

SaaS software as a service.

SEM search engine marketing.

SM social media.

SVD singular value decomposition.

WA web analytics.

WOM word-of-mouth.

WWW World Wide Web.

1 Introduction

More and more companies are getting involved with online business in order to get a share from this growing market. Besides only being present on the World Wide Web (WWW), they sell products and services right from their website, engage their customers using social media and advertise their businesses on online marketing platforms.

As a natural result of the increasing demand and usage of online business, related technologies have been significantly improved over the past two decades. Web analytics (WA) solutions have been used to analyze the customer behavior, and to get data to make sound marketing decisions. There are a variety of web analytics products which measure and report website usage in terms of performance metrics such as page views, visits, and conversions. These metrics and results can usually be broken down by several dimensions in order to make deeper analyses.

Social media (SM) is a newer and interesting way of communication for people, and for companies as well. Businesses use social media to publish content to reach their current and potential customers and to build another form of online presence. Another use of social media for companies is to enable people to reach them and share opinions with them. Since there are many social media platforms on the Internet, it is interesting to analyze brand recognition and equity using social media. As a result, some tools have been developed to measure and analyze social media data in a business point of view. Both the social media activities of companies and customer behavior and opinions are worth to analyze in order to support marketing, sales and public relations (PR) decisions.

Another interesting area for businesses is online marketing, specifically search engine marketing (SEM). Although it is a relatively new and evolving way of marketing, it already has a market share of several billion Euros. The possibility of targeting potential customers effectively makes this form of marketing very appealing. There are different advertising platforms and solutions for that. Using these platforms efficiently is an ongoing study by many parties. Companies are making use of some software tools and consultancy services to improve their online marketing campaigns in order to make the most out of their marketing budget.

There has been some research and developments in the above-mentioned fields, but social media is rather separated than the others. There is not much research on integration of social media with web analytics. The software implementations are also very immature. To the best of our knowledge, there is also no published work about making use of social media measurement data for online marketing purposes.

This work identifies the possibilities for bridging social media measurement with web analytics and search engine marketing. A methodology and framework has been suggested to retrieve and measure social media data and to align it with the already-existing web analytics data. This framework facilitates certain types of data analyses to support decision making for SEM.

Two types of analyses have been performed to utilize social media data for SEM. First analysis is in semantic level and uses some text analysis methods. It makes use of social media as a text corpus to suggest search engine marketing keywords. For this purpose, term extraction is performed first in order to find

candidate terms. Then, these terms are categorized and placed in an ontology which contains the domain knowledge of the field of interest. A novel methodology has been proposed for handling search engine marketing campaigns. It facilitates the creation and maintenance of campaigns by making use of ontology and structured ad groups.

As the second type of analysis, local interests on different products have been identified using social media mentions. This data has been compared with different data sources. The results show that, regional social media interests on different products are strongly correlated with product sales in corresponding regions. It is possible to make use of this method to make marketing decisions. One possibility is to focus a product-specific advertising campaign on an interested region to get higher return on investment (ROI).

The rest of this document is structured as follows: In Section 2, the main business areas of concern are introduced and some background information about them is given. Goals of the project and their motivations are stated in Section 3. Next, Section 4 lists the research problems encountered during this study and gives an overview of state-of-the-art in these fields. The selected solutions to these problems and other developed methods to solve subtasks of the project are explained in Section 5. Section 6 provides an overview of the results obtained and discusses about them. Finally, Section 7 lists our findings and contribution with this study and discusses about its implications. It also provides a comparison to the previous work, points out some limitations of the project and shows directions for future work.

2 Background Information

This project intends to fill the gap between the fields of web analytics, social media and online marketing. In this section, you can find some background information about these fields, which are the domain of this project.

2.1 Web Analytics

Web analytics can be described as the collection, measurement, analysis, and reporting of web usage data. This data is essential for web developers and marketers to understand the user behavior and interests. Statistics collected from web analytics solutions can be used to improve the website structure and content, which in turn increase the website users, sales and thus revenue.

2.1.1 Concepts

Web analytics tools measure the performance of a website in terms of some performance indicators. These indicators may include the number of visitors, number of page views, volume of sales, average time a visitor spends on a page, etc. In this context, these numeric values are called as *metrics*. Figure 1 shows the change of two metrics for a website over time.

Figure 1: Web analytics: Metrics

Website activities can be grouped and segmented by a number of criteria. These are called *dimensions* or *segments*. Some of the most typical dimensions are visited webpages, traffic source of the visitor, system properties of the user, search keywords, geographical location of the user, and time. Figure 2 shows some metrics where traffic sources are used as the dimension.

If a part of web usage information is relevant for a particular analysis instead of the complete information, the data can be sliced by some criteria. These criteria are called as *filters* and usually done by assigning a value to the dimensions described above.

2.1.2 Data Collection

In order to measure web usage, some data need to be collected. Different web analytics solutions use different techniques for this. One of the methods is to use server log files. Since most web servers log hyper-text transfer protocol (HTTP) requests, it is possible to get some web usage information by processing this file. Another method is to tag webpages. To do this, an invisible image or

	Source/Medium None	Visits ↓	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
1.	google / organic	**56,012**	1.59	00:01:03	84.69%	79.61%
2.	google.com.tr / referral	**18,577**	1.48	00:00:40	93.67%	66.16%
3.	(direct) / (none)	**18,566**	2.37	00:02:03	71.54%	66.93%
4.	cclub.metu.edu.tr / referral	**2,305**	3.86	00:03:23	58.74%	41.08%
5.	ceng.metu.edu.tr / referral	**1,881**	2.98	00:02:12	66.67%	52.37%
6.	facebook.com / referral	**993**	3.16	00:02:33	58.21%	46.83%
7.	google.com / referral	**802**	2.08	00:01:09	87.16%	59.98%
8.	bing / organic	**416**	1.27	00:00:29	95.43%	89.66%
9.	linux.org.tr / referral	**355**	4.67	00:03:13	77.46%	28.45%
10.	search / organic	**341**	1.43	00:00:46	94.43%	83.58%

Figure 2: Web analytics: Traffic sources as the dimension

a JavaScript code can be inserted into rendered webpages. This image/code makes a request and thus sends data to the server of the web analytics software. Both methods have some advantages and disadvantages with respect to each other. There are also other methods and some hybrid techniques to collect web usage data for web analytics solutions.

2.1.3 Products

There are several web analytics solutions available in the market. Some of them are freely available, while the others are sold commercially. There are two main categories for such solutions. The first group consists of on-premises software products, which are installed, configured and run on the site of customers who use the software. Therefore, website usage data is kept and displayed directly on the servers of the person or organization who owns the website. When this option is not practical for the user, second group of solutions can be preferred. In this group, web analytics software is distributed in software as a service (SaaS) form. In this scenario, pages of a website need to be configured in such a way that, website usage statistics are sent to the servers of the analytics software provider. Then, data can be seen and analyzed on a dashboard which is again hosted externally by the software provider. However, for some organizations where security and privacy are primary concerns (e.g. banks, military and government organizations) this type of solutions may not be desired. In this case, first group of solutions could be better alternatives.

Google Analytics[1] is by far the most widely used SaaS application for web analytics with a market share of 82%[2]. It is hosted by Google itself and is quite fast, robust and easy to use. It is also publicly available and provides sufficient functionality for most users. However, this functionality may be insufficient for some enterprises which require more advanced and customizable features. Or its data limits can be restrictive for high traffic websites. In such use cases, some

[1] `http://www.google.com/analytics/`
[2] `http://w3techs.com/technologies/overview/traffic_analysis/all`

commercial products by Omniture, Quantcast, Nielsen, etc. can be preferred by website owners.

2.2 Online Marketing

Online marketing is the activity of promoting and selling products and services over the Internet. There are many forms of online marketing and various platforms to facilitate these operations.

E-mail marketing is one of the most common forms of online marketing today. As its name implies, it is done by sending promotional e-mails to the current or potential clients. This category also includes spam mails.

Display advertising is another form of marketing where an advertisement material (text, image, banner, video etc.) is displayed to a user. The billing can be based on the number of displays, number of clicks, number of sales or on revenue share. It is possible for advertisers to publish their ads by connecting directly to websites who want to make money by displaying some ads. However, this solution can be inconvenient for the cases where many small advertisers and/or content providers are involved. In such cases, some advertising platforms can be used.

2.2.1 Online Marketing Platforms

There are different kinds of advertising platforms on the Internet. Big merchants create their own affiliate programs, so that small websites (affiliates) can display their ads to earn money. Amazon[3] is one of the most well-known example for such programs. On the other hand, major websites or content publishers can also create their own advertising programs, so that, small advertisers can register to publish their ads to some of the users of the publisher website. Publishers typically target the ads automatically based on several criteria like geolocation of the visitor, content of the page, user behavior, etc. Such programs are run by Facebook[4], MySpace[5], YouTube[6] and major search engines. The last type of advertising platforms brings publishers and advertisers together. These are called advertising networks. Examples of such networks include Google AdWords[7], DoubleClick[8] and Yahoo Publisher Network[9].

2.3 Search Engine Marketing

Search engines are used by most Internet users to find the relevant information on the World Wide Web. Normally, search results are listed and ranked according to their relevance with the search term (keyword) and the users. Besides these regular results, some search engines also display sponsored search results as in Figure 3. These results are basically online advertisements selected and ranked considering the relevance. In order to appear and have a higher rank in the sponsored search results, website owners and marketers create advertising

[3] https://affiliate-program.amazon.com/
[4] http://www.facebook.com/advertising/
[5] https://advertise.myspace.com/
[6] http://www.youtube.com/t/advertising_overview
[7] http://adwords.google.com/
[8] http://www.google.com/doubleclick/
[9] http://advertisingcentral.yahoo.com/publisher/index

campaigns and pay to search engine marketing platforms. This form of online advertisement is called search engine marketing (SEM). All major web search engines have their own advertisement platforms for SEM including Google, Yahoo! and Bing.

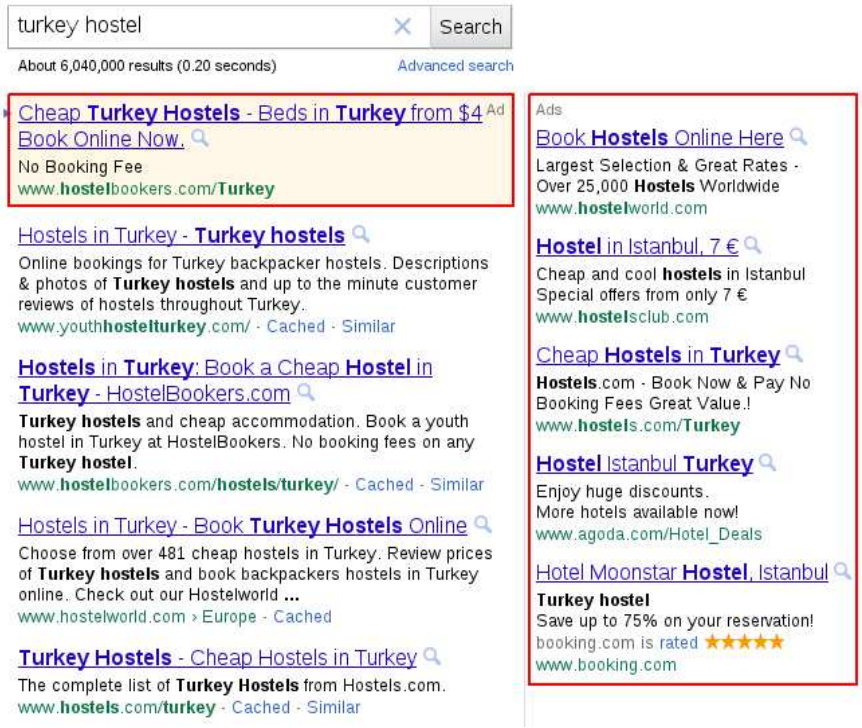


Figure 3: Sponsored ads in search engine results

2.3.1 How it works

In order to advertise on one of the search engines, one should sign up for an advertisement account in their SEM platforms. Then, it is possible to create some advertisement campaigns. A number of variables can be defined for these campaigns including budget, target location and language, time schedule, ad creatives, keywords and bids.

When a user performs a search on the search engine, it filters the advertisements where the ad parameters match with the search keyword and the attributes of the user. Then, an auction is performed among the qualified ads depending on their bids and relevance to the query. Successful ads are displayed along with the normal search results. Depending on the payment options, the advertisers pay the search engine based on the number of impressions (CPM) or the number of clicks (CPC).

Figure 4 shows the stakeholders of SEM and their interaction with the search engine.

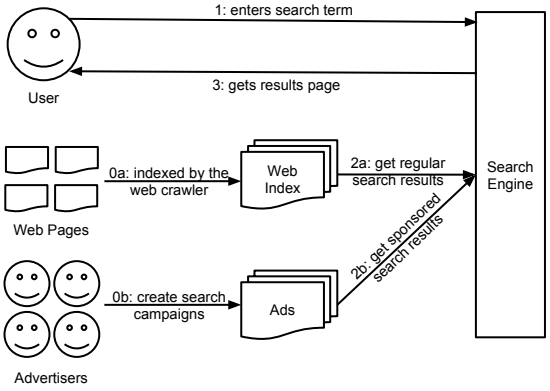


Figure 4: Search engine marketing stakeholders

2.3.2 Parameters

A person or organization typically has an advertising account on one or more search engine marketing platforms. The (simplified) structure of such an account is described below with a JSON-like syntax, where sets are defined in curly braces and square braces denote arrays/lists.

$$
\begin{aligned}
\text{Ad Account} &= [\text{Campaign}] \\
\text{Campaign} &= \{\text{Budget}, [\text{Ad Group}], \text{Schedule}, [\text{Location}], [\text{Language}]\} \\
\text{Ad Group} &= \{[\text{Ad Creative}], [\text{Keyword}]\} \\
\text{Ad Creative} &= \{\text{Title}, \text{Content}, \text{Target URL}\} \\
\text{Keyword} &= \{\text{Keyword Text}, \text{Maximum Bid}\}
\end{aligned}
$$

2.3.3 Campaign Management

In order to create a successful online advertising campaign, a few important steps should be taken. Today's common practice of campaign management is as follows: First, a couple of campaigns and associated text ads are created. Sets of keywords are selected manually or semi-automatically and associated to each ad group. Then, the value of a conversion (desirable action or purchase) and average conversion rate for the website is determined. Using this average revenue per click information, bids for keywords and/or ads are determined to increase the total revenue. Different ads and campaigns are compared and optimized to maximize the ROI. These steps are usually done either by an online marketing expert or a marketing campaign management tool.

There are a number of challenges involved in this procedure. First, determining relevant keywords is usually done intuitively with limited to no help of software. Second, bids need to be determined for each keyword individually. This is a very time consuming task if the number of keywords is high and additional help may be needed. Additionally, if the conversion rate is not known or estimated for a new keyword, it is not possible to determine a safe bid.

2.3.4 Gain of a Campaign

As it can be seen in Section 2.3.2, there are a number of parameters to define in an advertising campaign. They are all used to increase the gain obtained from the advertising campaign. There are already some formalization of the budget allocation and gain optimization problem in the literature [18, 10], mainly done by search marketing platform providers. These papers usually assume that most of the parameters are constant and the only variable is keyword bids. We follow another approach to be able to consider the effect of different parameters. We define the following terms in this domain:

K set of keywords

k a single keyword

$\mathbf{clicks}_k$ the number of clicks obtained from keyword k

$\mathbf{cpc}_k$ the average cost of a click for keyword k

$\mathbf{bid}_k$ the maximum cost per click (CPC) (bid) for keyword k

$\mathbf{rev}_k$ the average revenue obtained from a successful conversion/sale for k

$\mathbf{conv}_k$ conversion rate of customers visited by searching k

According to these definitions, we can calculate the gain of an online search engine marketing campaign as:

$$\text{gain} = \sum_{k \in K} \text{clicks}_k \cdot \frac{\text{rev}_k \cdot \text{conv}_k}{\text{cpc}_k}$$

Since the number of clicks and CPC depends on the budget and bid on keywords, we can modify the formula as:

$$\text{gain} = \sum_{k \in K} F(\text{budget}, \text{bid}_k) \cdot \text{rev}_k \cdot \text{conv}_k \tag{1}$$

2.4 Social Media

Social media can be described as the collection of Internet-based applications which allow communication and content publication and sharing among their users. Social media have many different forms, including forums, blogs, wiki's; microblogging, social networking, social bookmarking, photo and video sharing applications. Some of the most commonly used social media platforms are Facebook, Twitter, YouTube, Flickr and Blogger.

2.4.1 Social Media Monitoring

With the increasing popularity of Web 2.0 and social media platforms, tremendous amount of user-generated data became available. This data created an opportunity to monitor and measure user activity and opinion for digital marketing and public relations analysis purposes. The importance of social media measurement for organizations is two-folds. First, they can monitor the success of their social media activities by tracking their own social media accounts. Second, they can analyze public awareness and sentiment of their brands and products.

2.4.2 Sentiment Analysis

For the second goal mentioned above, it might be necessary to detect the sentimental polarity of user posts on social media in order to analyze their opinion. Although it can be done manually for a limited number of messages, it becomes too labor-intensive when large volumes of data need to be processed. Therefore, some software can be developed and used to facilitate this operation. This application of identifying and extracting subjective information in textual sources is called *opinion mining* or *sentiment analysis*.

2.4.3 Monitoring Tools

Several entry-level and enterprise platforms are created in order to monitor off-site user activities by monitoring social media. The companies which offer such solutions and their tools include Twitter Sentiment, Social Mention, TweetFeel, Radian6, Sysomos, Altarian, Autonomy, RapidSentilyzer and Buzzcapture. Some of these tools are free to use while the others are commercially available. In Figure 5, one of these tools can be seen in action.

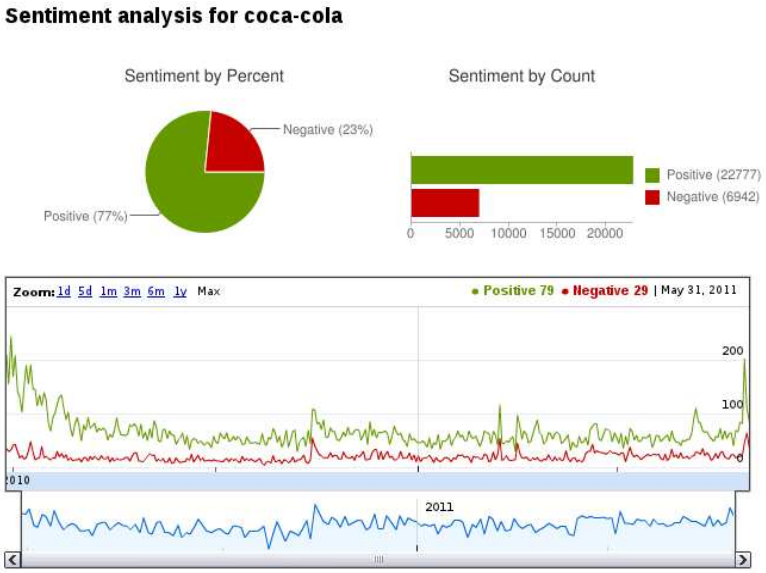


Figure 5: Social Media Monitoring

Most of the current social media monitoring tools have features to track social media sources, most notably Twitter. Some of them can aggregate data retrieved from different sources. Similarly, some tools support sentiment analysis, customer engagement options, configurable alerts, etc. On the contrary to the web analytics tools, the features and standards are not well-established and heavily depend on the particular tool. Even though measuring social media activities and web-site usage activities share very similar characteristics, integration of social media monitoring and web analytics tools are quite limited.

3 Motivation and Goals

The domains explained in the previous section have being developed continuously. However, as these fields are becoming more mature, the gap between them is growing. Our aim is to identify the possibilities and provide solutions in order to bridge this gap.

Consequently, two main goals have been defined for this project. First, a framework will be built for integrating social media measurement data with web analytics data. Second, the combined data will be analyzed for making SEM decisions. Figure 6 illustrates the relationship between these three fields and how our goals help bridging them. The details of these goals and their motivations can be found in the corresponding sections.

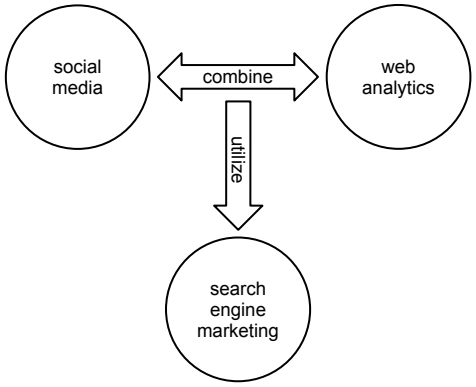


Figure 6: Relationship between SM, WA and SEM

3.1 Aligning Social Media with Web Analytics

There are several web analytics and social media measurement tools available in the market. In web analytics domain, there is an established design for such frameworks. In this scheme, user activities on websites are observed and this information is stored. Then, aggregated results are presented to the website owner in terms of metrics and dimensions. Social media monitoring domain, on the other hand, is relatively new and most tools follow a different design.

Furthermore, these closely related tasks are usually not managed together. Even though some web analytics tools offer limited social media monitoring features, these are very immature. There is a strong need to standardize social media measurement activities and integrate them with existing web analytics solutions.

Being able to combine web usage and social media statistics together should enable new types of reports and analyses. It will be possible to see the correlation and effect of these two domains on each other. For instance, one can see the effect of a social media marketing campaign on website visits and sales. Another possibility is to see the popularity of different products on social media and comparing this information with sales numbers to measure the company's ability to capture the market demand.

Therefore, one of the goals of this project is to design and implement a framework which collects data from social media and aligns it with web analytics data to enable new types of analyses.

3.2 Utilization for Search Engine Marketing

Search marketing campaigns, which are used to advertise a website on search engines, are usually created manually and improved by analyzing campaign statistics and web analytics data. Although social media data can be also useful to manage advertising campaigns, there is no published work which considers this possibility, to the best of our knowledge.

It has been stated that, the suggested social media measurement framework in this project enables new types of analyses to utilize social media data. In order to verify this claim and to explore the usability of social media data on search engine marketing domain, some methods and experiments need to be designed. The goal of these methods is to utilize the data retrieved from social media in order to create and improve search engine advertising campaigns.

Considering the Formula 1 in page 8, we have tried to identify possible methods, in which social media data can be used to improve advertising gain. It can be observed that, the average revenue of a keyword, which appears as rev_k in the formula, mostly depends on the product cost and price, and is not much related to the campaign parameters. That is why; we have focused on other parameters to improve the gain. At the end, we have identified two methods which can be useful to improve the efficiency of online marketing campaigns: using social media data to generate SEM keywords; and capturing local interests to focus campaigns more effectively. These methods are described in the following sections.

3.2.1 Keyword Generation

In order to create advertising campaigns, bids need to be put on search keywords. For this reason, these keywords need to be determined. This parameter appears as set K in Formula 1. There are two important criteria to determine keywords.

First, the keywords should be relevant to the advertisement and the target website. Otherwise, even if the number of impressions can be high for the advertisement, click-through rate (CTR) and conversion rates could be very low, which makes the campaign inefficient.

Second, it is important to find cheaper keywords. It has been observed that the popularity of the keywords follow a long tail distribution [1] as in Figure 7. This means that, a small portion of search terms are responsible of a significant amount of traffic. On the other hand, a larger number of terms with low search volume cumulatively make up also a significant share of the total traffic. Since highly popular keywords are generally more expensive in search auctions, the goal of keyword selection task should be to find a big number of relevant keywords in the long-tail of the keyword popularity distribution.

In order to understand the importance of finding cheaper keywords, consider the following example: An insurance company wants to increase the number of their website visitors and sell more insurance policies. They create a SEM campaign and advertise on “insurance” keyword. Since it is a popular keyword, they have to bid at least 5€, so that their bid wins the search auction. If

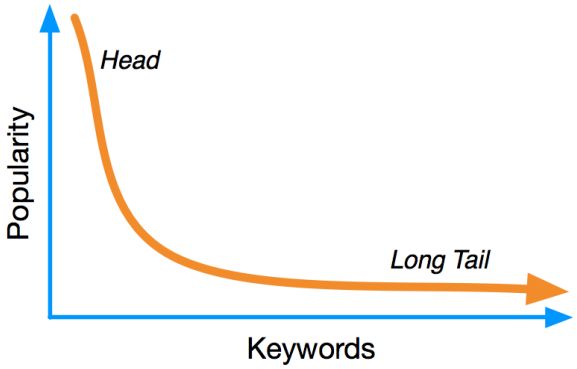


Figure 7: Long-tail distribution of keywords

they have 500€ as marketing budget, they can get only about 100 visitors. However, they can choose to use a number of less popular keywords such as "car insurance", "health insurance" and "life insurance". The average cost for these keywords is 2€. Therefore, they can get 250 visitors instead of 100 with the same budget. If they find less popular keywords with the same meaning, they can increase the efficiency of their campaign even further. For example, they can use "auto insurance" and "vehicle insurance" keywords, which are cheaper than "car insurance", but have the same meaning and the same amount of traffic when summed up.

Currently, several methods are used by marketers to determine the keywords of a campaign. Most common approach is manual selection which is done by a domain expert intuitively. They can also use some tools provided by search engine marketing platforms such as Google AdWords[10]. Such tools provide lists of categorized keywords, find semantically related keywords given initial keywords, and extract possible keywords using a webpage content. Another useful tool would be one which gives statistics about keywords and number of searches by location such as Google Trends[11] and Google Search Insights[12].

We have identified and compared several methods to find SEM keywords. The details can be found in Appendix C in page 48. In order to make use of our framework and see the usability of social media data in search engine marketing, we try to focus on methods which can use social media data. So, news articles, blog posts and online buzz can be used to generate more keywords which are related to the campaign. The proposed solution for this goal can be found in Section 5.3.

3.2.2 Detecting Local Interests

Another important parameter in Formula 1 is conv_k. Increasing the conversion rate obviously contributes to the gain of the campaign. However, this single parameter depends on many factors such as the relevancy of the advertisement to the target page and product, competitiveness of the product offer on the

[10] AdWords Keyword Tool - `https://adwords.google.com/select/KeywordToolExternal`
[11] `www.google.com/trends`
[12] `www.google.com/insights/search`

website, ease of use of the website, etc. It also depends on the willingness of the users to buy the corresponding products. Besides selecting relevant keywords, interested users may be targeted by focusing on certain geographical locations.

Our intuition and hypothesis is that, the users living in a certain geographical region can be more interested into a particular product or product family compared to the ones living in other locations. If this hypothesis can be verified, targeting the campaign on such geographical regions where more people are interested into a product p and a related keyword k would result in a higher conversion rate for k (conv_k) and thus increases the overall gain.

Consider the following situation: A car manufacturer produces different models. The company wants to enter Turkish automobile market, where they did not exist before. They want to advertise on search engines to get a market share. They advertise their models in all the country. Their campaign leads many visitors, but they make only a few sales. By making a local popularity analysis, they find out that, different types of cars are popular in different cities because of social, economical and geographical differences between the cities. After the analysis, they split their advertising campaigns into different product groups and they focus each campaign on more interested regions. This doesn't change their cost per click. However, by targeting customers with more relevant products, they increase their conversion rate, sales and revenue with the same budget.

It is possible to make such modifications on the advertisement strategy, since most search engine marketing platforms offer the possibility of selecting geographical regions for advertising campaigns. Then, budget and bids of the targeted campaigns can be increased compared to the corresponding unfocused campaign.

Finding interested regions can be done by creating a campaign initially, and then analyzing the conversion rates for location-product pairs in campaign statistics and web analytics data. However, there are some shortcomings of this approach. First, the website, campaign and web analytics application should already be running in order to have statistics for conversion rate. Second, the products which will be advertised should already be on the website. It is not possible to make such an analysis for new or upcoming products. It is also not possible to take an action before the conversion rate changes because future changes of conv_k cannot be predicted.

Two alternative indicators can be used to specify target location of a campaign. Assuming that search interests and purchase interests are correlated, web search statistics can be a useful source to determine campaign targets. If a particular city or region performs more searches about a product compared to the other regions, conversion rate can be higher for that region. Thus, bids on CPC can be increased. Likewise, mentions and sentiments in social media can be used to determine which regions are interested in which products or services. In this project, we will focus on the latter.

4 Research Problems

In this section, some research problems encountered in the scope of this project are introduced. Related work and the state-of-the-art in these fields are explained.

First research topic, sentiment analysis, is needed to improve social media data with sentimental polarity information. This data is useful for business purposes such as brand analysis. Second topic, keyword generation, is one of the goals of the project as it is explained in Section 3.2.1. Finding relevant keywords are important for creating a successful advertising campaign. Ontology learning, which is introduced as the third research problem here, has been faced while creating keywords using social media data. Since some domain knowledge was required to create such campaigns, ontology learning is used as a method to obtain such knowledge. Last of all, as one of the main goals of this project, search engine marketing campaigns are tried to be improved using social media data. That is why, campaign optimization problem and previous work on this field are described in this section.

The relationships of these research problems, the goals of the project and components of the developed framework are shown in Figure 8.

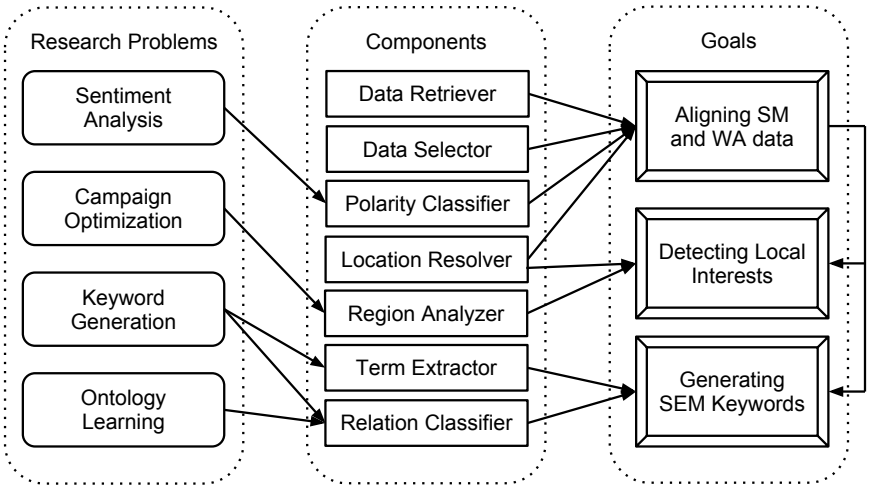


Figure 8: Relations between problems, components and goals

4.1 Sentiment Analysis

As it is already discussed in the previous section, *sentiment analysis* or *opinion mining* can be described as the automatic analysis and classification of opinion, sentiment and/or subjectivity in text. The basic task here is to classify the *polarity* of a given text at the document, sentence or feature/aspect level.

At this point, it is also useful to define *classification*. It is a machine learning task of inferring classes (types) of objects using training data. The training data contains set of examples where input objects are associated with classes. Classification algorithms use this knowledge to predict the class types for input

objects for unknown items. Classification is a specific case of *supervised learning* where the output values are discrete (set of classes).

Sentiment analysis is a special form of supervised learning where input objects are pieces of text like sentences, messages or document and outputs are values which indicate the sentimental polarity of the inputs. It is also possible to determine a set of discrete output range of classes such as positive, negative and neutral.

Since most classification algorithms cannot process text inputs directly, it is necessary to preprocess the text to transform it into a feature vector. A feature is a certain characteristic of the input object. It may, for instance, denote the presence of a particular word. The features need to be predetermined for classification to make sure all the input objects have the same number and type of values. Feature selection is therefore a crucial part of sentiment analysis along with classifier and parameter selection.

4.1.1 Related Work

Computer science community has been doing research on this relatively new field for a decade. Pang and Lee [23] is a good survey which describes the basics of the field, its subproblems, methods, applications areas, challenges and available resources.

The accuracy of different machine learning methods with various features for sentiment analysis problem have been evaluated by Pang et al. [24].

Some research papers have a focus on sentiment analysis on data obtained from a particular source. The most common example is Twitter. Go et al. [13], Pak and Paroubek [22] have both worked on opinion mining on Twitter domain, considering the properties and limitations of this environment.

Specific cases of sentiment analysis have also been investigated by various researchers. Jindal and Liu [16], Ganapathibhotla and Liu [12] focus on evaluating the sentimental polarity of texts on comperative sentences such as "I prefer Coke rather than Pepsi". Similarly, Narayanan et al. [20] investigates the sentiment on conditional sentences.

Various aspects of the field have also been investigated. For example, while the most work is done on classifying sentiment in document-level, Ding et al. [8] focuses on entity discovery to identify which entity is being discussed in the text. On the other hand, Devillers et al. [7] considers different kind of human emotions like anger and fear, rather than making an analysis only one dimension with negative-positive scale.

Apart from the sentiment analysis techniques, the application areas of this task have been studied by many scholars. Since the predicting the number of product sales is very important for companies, the utilization of sentiment analysis techniques for sales predictions have been investigated by Reijden and Koppius [26], Gruhl et al. [15]. Another similar prediction task has been performed by Krauss et al. [17] to identify the importance of word-of-mouth (WOM) on product success on the domain of movies. OConnor et al. [21] shows the correlation between sentiment analysis results and traditional polls and discusses that, opinion mining can be a substitute or supplement for traditional polling. Last of all, Bollen et al. [3] applies sentiment analysis for stock management purposes and suggests that Twitter data is and moods are useful to predict the stock market.

4.1.2 Resources

Different tools and resources are available for sentiment analysis tasks. The first group of resources includes open source machine learning tools. These tools can be used to implement several machine learning tasks in order to perform sentiment analysis. WEKA[13] is one of the most mature and commonly used libraries and has tools for preprocessing, classification, clustering and visualization. Apache Mahout[14] is a similar library, which is introduced more recently and focuses on scalable machine learning algorithms.

Some other tools can also be helpful to solve partial problems. For natural language processing (NLP) related subtasks, some NLP libraries can be used. These include GATE[15] and OpenNLP[16]. Specifically, if part-of-speech (POS) tags need to be used as features for classification, POS taggers can be used. Similarly, word stemmers can be used to convert inflected words to their basic forms for easy processing. Task-specific tools exist for this kind of purposes. For example, Stanford Log-linear Part-Of-Speech Tagger[17] and Snowball[18] can be used for POS-tagging and word stemming, respectively.

Some lexical resources may be useful for text preprocessing and opinion mining. One may want to eliminate stop words in a bag-of-words classification model. Several stop word lists are available online for this purpose. Also some of the NLP and machine learning tools mentioned above contain their built-in stop word lists for some languages. Another useful lexical resource is SensiWordNet[19]. It assigns each synset of WordNet, which is a set of words with the same meaning, three sentiment scores: positivity, negativity and objectivity. This sentiment scores can be used to compute the sentimental polarity of a whole sentence or document without applying a machine learning algorithm.

Although not many tools and services are available for sentiment analysis directly, there are still some resources. OpinionFinder[20] is a sentiment analysis library, which can be used for research purposes. On the other hand, TwitterSentiment[21] project provides a free web API for text sentiment classification.

4.2 Query Expansion / Keyword Generation

Query expansion is the process of improving an initial query for information retrieval (IR) operations. Starting with the seed query, the procedure creates a more complex query or a number of basic queries which are intended to improve the performance of the operation. Some performance metrics are precision and recall, similarly to many IR tasks. The aim is usually to increase *recall* (the fraction of the relevant results returned among all relevant items) while keeping *precision* (the fraction of the relevant items among returned results) as high as possible.

Keyword generation for search marketing is somewhat similar in nature.

[13] http://www.cs.waikato.ac.nz/ml/weka/
[14] http://mahout.apache.org/
[15] http://gate.ac.uk/
[16] http://incubator.apache.org/opennlp/
[17] http://nlp.stanford.edu/software/tagger.shtml
[18] http://snowball.tartarus.org/
[19] http://sentiwordnet.isti.cnr.it/
[20] http://www.cs.pitt.edu/mpqa/opinionfinderrelease/
[21] http://twittersentiment.appspot.com/

This time, precision and recall metrics are based on the number of reached users, as opposed to the number of returned documents. By generating more keywords, marketers aim to reach more users who are interested into the company or product (increasing recall). But it is important to use only relevant keywords to reach only to interested users (retaining high precision), because displaying advertisements to uninterested users increases the advertisement costs while not contributing to the revenue.

4.2.1 Related Work

There is limited research in the area of keyword generation for search engine marketing. Some of the current work including Fuxman et al. [11], Yih et al. [34] utilizes search engine query logs to suggest new keywords based on an initial search term. However, this data source is not available to public. Abhishek and Hosanagar [1] takes another approach to use term similarity for keyword generation. They calculate term similarity by building a corpus by crawling and indexing relevant web pages.

4.3 Ontology Learning

An *ontology* is an explicit specification of conceptualization [14]. It is used to formalize the knowledge in a domain of interest. The knowledge is usually represented as a set of concepts and the relationships between them. This formalized representation of the knowledge can be used by computers for several purposes and applications.

There are several knowledge representation language standards which specify how to define an ontology. Examples include RDF[22] and OWL[23]. On the other hand, several ontologies have been published by researches. Each of these ontologies is used to describe knowledge in a particular domain. For example, WordNet[24] is a lexical database of English language which lists English words and their relationships. Similarly, there are several other ontologies in domains such as biology, medicine, geography, web and human relationships.

Ontologies can be built by domain experts manually. Alternatively, methods can be developed to support automatic or semi-automatic ontology development. It is referred to as *ontology learning* [5]. It is concerned with automatic knowledge acquisition, mostly from a text corpus. It is closely related to NLP, machine learning and information retrieval fields. A very basic overview of ontology learning is given in Figure 9.

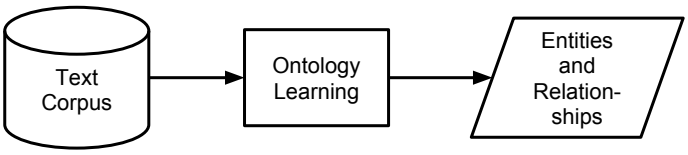


Figure 9: Ontology Learning Overview

[22]http://www.w3.org/TR/rdf-syntax-grammar/
[23]http://www.w3.org/TR/owl-overview/
[24]http://wordnet.princeton.edu/

4.3.1 Related Work

A survey which brings together some selected papers on ontology learning is Buitelaar et al. [5]. They classify the sub tasks of the area from easiest to hardest as: extracting terms, finding synonyms, extracting concepts, detect concept hierarchies, finding concept relationships and extracting rules.

The problem of finding synonyms and semantically similar words has been addressed to many researchers. A part of a survey about vector space models has been dedicated to word similarities and synonyms [31, sec 2.2]. They bring together the significant works about this problem which has published before. Additionally, some research activities focus on different methods of finding synonyms. Turney [29] suggests a method to use a web search engine to estimate the semantic similarity between a pair of words. They test their method on TOEFL questions to find the synonyms among four alternative word pairs. Their findings show that they perform better than an average student. Baroni and Bisi [2] takes a similar approach and compares this method with other alternative techniques available.

Another group of papers focus on extracting other kind of word relations. Turney [32] suggests a method to analyze text patterns to detect the type of relationships between word pairs. The technique is based on their *latent relation hypothesis*, which basically suggests that, if two pair of words share a text pattern, it is more likely that they have a similar relationship. This is claimed to be a uniform approach to detect many types of word relationships. Nakov and Hearst [19] also takes a linguistic-based approach to characterize the relation between a pair of words.

There is a significant effort made in ontology learning in biomedical medical. Papers such as Rindflesch et al. [27], Pustejovsky et al. [25], Vintar et al. [33], Buitelaar et al. [4] and Ciaramita et al. [6] are all focused on extracting domain knowledge such as various relationships between medicines, diseases, genes and cells. They commonly use medial publications database as their text corpus.

4.4 Campaign Optimization

The basics of search engine marketing have been explained in Section 2.3. In this model, advertisers create advertising campaigns to attract more users to their website. The aim is usually to increase the revenue of their company by spending the advertisement budget efficiently. The process of improving advertising campaigns is called as *campaign optimization*.

The optimization is done by changing the available parameters in the campaigns such as keywords, bids, target audience, ad contents, etc. Since many parameters are involved and the results can mostly not be predicted beforehand, it is a complex procedure which should be handled carefully. Because, changing parameters improperly may result in less user clicks and conversions and thus a loss of money.

4.4.1 Related Work

Edelman et al. [9] describes the basics of the principles of the search engine marketing. It emphasizes on generalized second-price auction model, which is being used by most search engine advertising platforms today. A clear understanding of this model is necessary to create more efficient advertising campaigns.

There is limited research in the area of campaign optimization. Most of the published research focus on the so-called budget optimization problem. Common definition of this problem assumes a fixed advertising budget and a static set of keywords to find the optimal distribution of budget and bids on the keywords. Rusmevichientong and Williamson [28] suggests an adaptive method to improve the budget distribution to satisfy a near-optimal profit. Feldman et al. [10] suggests that a randomized uniform bidding strategy gets a close number of clicks to the maximum clicks possible. Muthukrishnan et al. [18], on the other hand, focuses on both predicting the outcome of a budget distribution, as well as maximizing it. They suggest three scholastic models to answer these questions. Unfortunately, a little, if any, published work exists which also consider other campaign parameters such as target region, ad contents and campaign schedule.

5 Methods

Considering the project goals given in Section 3, and the necessary subtasks to be performed, some methods have been developed and used for this project. First, the design of the framework and integration methods are explained. Then, the sentiment classification method used in polarity classifier component of the framework is described. In Sections 5.3 and 5.4, the proposed solutions for SEM optimization are given. The details can be found in following sections.

5.1 Design of the Framework

One of the goals of this project is to align web analytics and social media data. In order to make a logical alignment with web analytics data, social media measurement data should be aggregated in a similar way that web analytics tools aggregate webpage usage activities. To be more specific, aggregated values should be represented as metrics, and those results should be broken down and filtered by dimensions. We can define metrics and dimensions in social media very similar to web analytics. Table 1 shows a very basic alignment between the two.

		Web Analytics	Social Media
metrics	atomic events	page views	posts
	bigger events	visits	conversations
	unique	visitors	users
	success events	conversions	positive messages
	success rate	conversion rate	percent positive
dimensions	date/time	aligned directly	aligned directly
	location	aligned directly	aligned directly
	medium	direct, organic, referral	twitter, news, blog
	source	website	publisher[a]
	keyword	aligned directly	aligned directly

[a] username is used for social media platforms like Twitter whereas domain name of the website domain is used for blogs and news items

Table 1: Aligning Social Media and Web Analytics Data

Once a logical alignment is made between WA and SM data, it is possible to combine them by making integration with web analytics tools. Figure 10 illustrates the position of the suggested framework with respect to WA solutions.

The proposed social media measurement and web analytics integration framework needs to be interacting with many external parties. The diagram given in Figure 11, shows these external parties and the data flow between them. As it can be seen, social media data needs to be retrieved from their sources. We have decided to focus on Twitter, blog and news items for practical reasons. The exact process of retrieving data from these sources is explained in Implementation Details section. The framework also needs to be communicating with web analytics software for integration purposes. Additionally, an external API

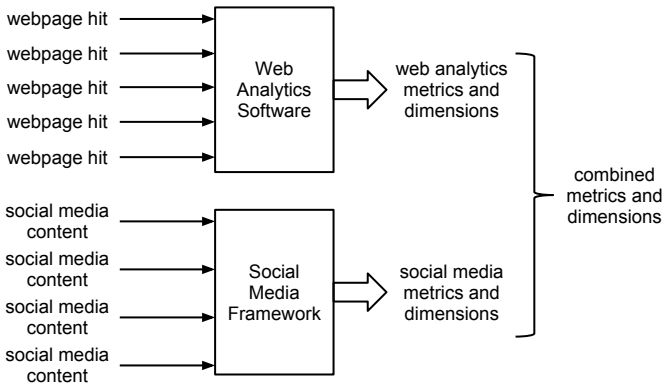


Figure 10: Combining Web Analytics and Social Media

is used to process geolocation information. Similarly, the details for these can be found in Appendix A.

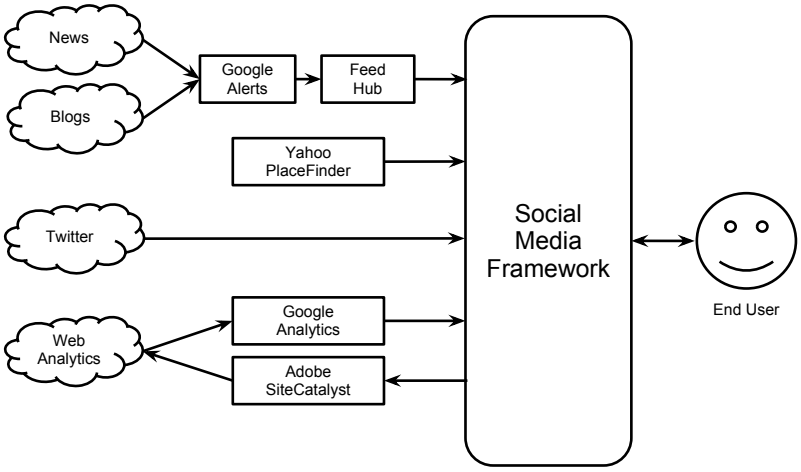


Figure 11: Architecture Diagram Level-0

The designed system is composed of a number of components as it can be seen in Figure 12. Data retriever component is connected to the Twitter, news and blog data sources. It retrieves relevant data from these sources and stores in the database in proper format. Data preprocessor component gets recently retrieved data and associates them with necessary metadata such as sentimental polarity and location information. Data selector component is used to access, filter, aggregate and format the data for the use of other components. WA integration component is responsible of importing web analytics data and exporting social media data to WA tools. Data analyzer contains some tools which are used to perform some data analysis for SEM purposes. Last of all, user interface component provides a basic web interface to select and display

the data in the system. The implementation details for these components can be found in Appendix A.

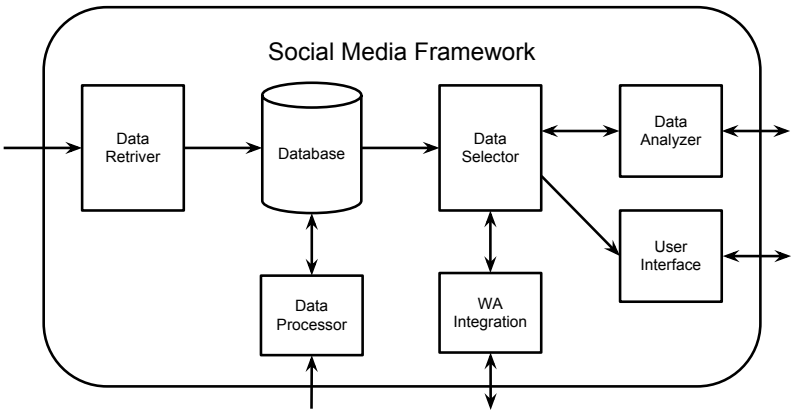


Figure 12: Architecture Diagram Level-1

5.2 Sentiment Classification

In order to achieve some of the goals of the project, sentimental polarities of messages need to be understood. As it can be seen in Section 4.1, sentiment analysis is a big and challenging field which is out of the scope of this project. Therefore, a basic technique has been selected for this project to classify social media messages as "positive" or "negative". Even though the outcome and accuracy is not perfect for individual messages, in aggregated level, it gives an intuition about the general feeling of users. Below, a step-by-step description of this method can be found:

Training and Test sets Training and test sets are needed in order to apply machine learning methods for classifying sentiments. Since there is no public manually-tagged data for social media messages, new training and test sets have been created similar to the method mentioned in [13].

According to this method, status messages with emotions (smileys) have been retrieved from Twitter. Those messages with smileys (e.g. ":)" ":D" ":-)") are assumed to have a positive sentiment and those with frownies (e.g. ":(", ":'(") are assumed to contain negative sentiments. Note that, this assumption changes the class definitions. The algorithms will try to guess the likelihood of a message to have a smiley and frowny instead of positive or negative sentiment.

In order to build the training set, 100,000 messages with positive emotions and 100,000 messages with negative emotions have been retrieved both for English and Dutch languages. Test sets have also been created with the exact same way. In order to eliminate the duplicates in the two sets, data has been collected in different time slots for training and test sets.

Extracting Features Machine learning algorithms need explicitly defined features to perform classification. In our implementation, the content of each

message has been tokenized and unigrams (words) have been selected as the features. Part of speech tagging has not been done because previous work [24, 13] reports that, it has little to no contribution to the accuracy.

Some text preprocessing is done during feature extraction. URL's, Hashtag's and mentions are removed from the message contents. Smilies and frownies have been replaced by "POSITIVEEMOTION" and "NEGATIVEEMOTION" equivalence classes. Out of all features generated, the most frequent n words have been selected to simplify the model.

Classifying Instances As the classification algorithm, Naive Bayes classifier has been chosen because of its simplicity and promising results. The classifier models have been built using the training sets for English and Dutch languages. These models are used for classifying future messages.

5.3 Keyword Generation

Creating advertising campaigns for search engines is usually done by domain and marketing experts with little to no help of automated tools. However, as the number of ad groups and keywords increase, it becomes much harder to manage such campaigns. They also become more likely to have mistakes.

Considering the general structure of SEM campaigns and observing some commercial settings, we came up with a suggested methodology for creating and managing advertising campaigns. This technique consists of the following steps:

1. Find relevant words for the context (see Section 5.3.1)
2. Build an ontology containing these terms (see Section 5.3.2)
3. Create ad groups and write advertisement texts
4. Express keywords in terms of patterns
5. Use the ontology and patterns to generate all keywords
6. Post-process keywords to enable/disable them and determine bids.

As it can be seen, social media data can be used as a corpus to facilitate the tasks stated in steps 1 and 2. Third and fourth steps need to be done manually by marketing experts. The last two steps can be done automatically by developing the necessary tools. This technique not only facilitates the creation of a campaign, but also makes it much easier to manage existing campaigns.

In our approach, the social media measurement framework is used to collect user-published content in the domain of interest. For instance, if the framework is used to support marketing activities of a hospital, those social media posts, which are related to healthcare domain, are collected. This corpus is used to obtain some structured domain knowledge (ontology) about the field. The ontology contains products and services of the company, and the relationships between them. It is also expected to contain alternative words and related terms. Then, given the ad group structure and keyword patterns which refer to the ontology entities, the exact set of keywords can be generated automatically as illustrated in Figure 13.

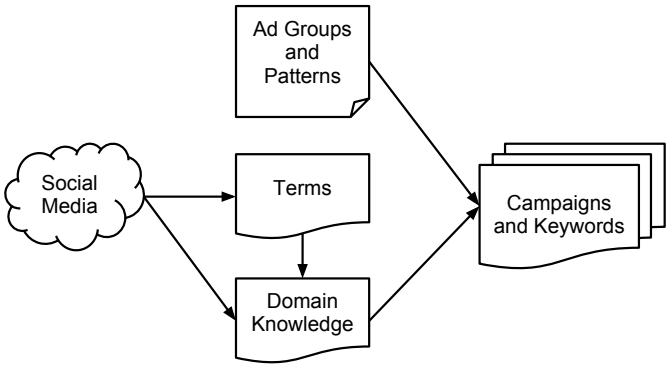


Figure 13: Keyword generation

For the first two steps of these tasks, namely term extraction and ontology learning, some automated methods have been developed as explained in the following sections.

5.3.1 Term Extraction

User posts retrieved from social media are used to extract context-specific terms. We apply a basic heuristics for this task where two datasets are used. The first dataset contains all posts published in a particular time range. The second dataset, on the other hand, is built as a result of a query. Therefore, it is expected to contain more context-specific terms with higher frequency.

According to this hypothesis, a simple heuristics is developed. For every word w in the datasets, its frequency in the general dataset (f_g) is compared to the frequency in the context specific dataset (f_s). If f_s is greater than a linear function of f_g (see Formula 2), the term is considered to be specific to the context. For an application of this method and the obtained terms, see Section 6.1.3.

$$f_s \geq m \cdot f_g + n \tag{2}$$

$$f_s \geq 5 \cdot f_g + 0.005\% \tag{3}$$

5.3.2 Ontology Learning

The terms extracted with the method described in the previous section are useful. However, most of the keywords used in search engine marketing consist of more than one word. In order to combine these words to create meaningful search keywords, some domain knowledge is required. This knowledge can be semi-automatically developed as ontology.

An ontology consists of a set of concepts and their relationships in a domain. We have developed a method to extend an existing ontology by adapting a technique called as latent relational analysis (LRA). LRA is used to detect the type of relationship between a pair of words. It is based on *latent relational hypothesis*, which suggests that pairs of words that co-occur in similar patterns tend to have similar semantic relations [30].

In this method, for each known relationship, a data corpus is searched where the two words in the relationship co-occur in close positions. All patterns containing any of these pairs are extracted. A typical pattern consists of 3-5 words, for instance, "w_1 of w_2 is *", "w_1 has * w_2" and "w_1 released new w_2". As it can be seen, wildcard patterns are used to increase the number of common patterns.

After finding all patterns, a matrix is generated where the rows correspond to word pairs and the columns correspond to patterns. Each cell contains the number of instances where the pair of terms occurs in the corresponding pattern. We also include the type of relationship as another feature (column) for this matrix. Using this setting, which can be seen in Table 2, it is possible to predict the type of relationship for a new pair of words as a classification problem.

	pattern #1	pattern #2	...	type
pair #1	1	3	...	relationshipType1
pair #2	7	0	...	relationshipType1
pair #3	2	14	...	relationshipType2
...			...	...
pair #n	...	...	...	relationshipTypeM
new pair	...	...	...	?

Table 2: Sample problem setting for relationship classification

By creating an initial ontology definition, and expanding it by supervised learning methods as described above, the necessary domain knowledge for keyword generation will be obtained. This knowledge can then be combined by ad groups and keyword patterns to generate actual SEM keywords.

5.4 Local Product Popularity Analysis

In Section 3.2.2, we have introduced a hypothesis which suggests that, different products are popular in different regions. In order to verify this claim and make use of it, a method has been developed to analyze local interests on different products.

Our solution uses the social media framework to compute regional interests. For each social media message related to one of the keywords, the geographical location of the author is found. Then, for each region-keyword pair, aggregated message frequency and sentiment are calculated. If keywords represent products and geographical regions are selected as cities, the results show the interests of each city in each product.

If our hypothesis can be verified, these results can be used to create geographically focused advertising campaigns based on the generic initial campaigns as shown in Figure 14.

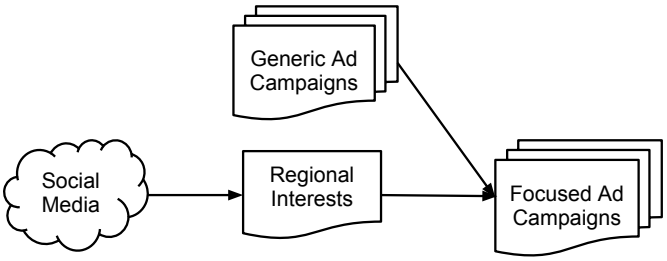


Figure 14: Focusing on geographical locations

6 Results and Discussion

The main measurable results of the project are given in this section. First, the accuracy of the sub-tasks of the project have been explored. Then, a case study has been made using the actual data from a company, and the results are given in a business point of view.

6.1 Accuracy of the Tasks

6.1.1 Sentiment Classification

The accuracy of our Naive Bayes classifier used for detecting sentiment polarity has been analyzed. Using the training sets and test sets created, as explained in Section 5.2, the confusion matrix in Table 3 has been obtained. The accuracy in terms of the percentage of correctly classified instances and Kappa coefficient are given in Table 4.

		Classified as		
Language	Class	Positive	Negative	Total
English	Real Positive	7442	2558	10000
	Real Negative	2017	7983	10000
	Total	9459	10541	20000
Dutch	Real Positive	6819	3181	10000
	Real Negative	1998	8002	10000
	Total	8817	11183	20000

Table 3: Confusion matrix of sentiment classification

Note that, there is a slight difference of positive/negative semantics of the classified messages and actual sentiment of the messages. Because, the test set has not been manually annotated as positive or negative. Instead, messages with happy or sad emotions are assumed to contain positive and negative sentiments, respectively.

	English	Dutch	Baseline
Accuracy	77.13%	74.11%	50.00%
Kappa	0.543	0.482	0.000

Table 4: Accuracy of sentiment classification

It is very hard to achieve 100% accuracy in automated sentiment analysis, because even human judges can disagree about the sentiment of a message. Most state-of-the-art studies[24, 13, 23] report accuracy levels between 80%–85%. On the other hand, a random classification on our data set would result in 50% accuracy, because the numbers of negative and positive messages are the same. Comparing these upper and lower baselines, our method performs

reasonably well. Although it could be possible to have a higher accuracy by improving this method, considering the simplicity of our approach and the fact that, sentiment classification is not a main goal of the project, these results are considered to be satisfying.

It is also important to note that, in most kind of analyses made with social media data only considers aggregated sentiment. In other words, it is only important to know how many of the messages are positive, instead of knowing exactly which messages are positive. Since false negative and false positive errors will most likely cancel out each other, aggregated accuracy is expected to be higher.

6.1.2 Resolving Location

One of the questions regarding to the framework is the correctness of the location resolver module, which detects the country and city of the author of a tweet. In Twitter, some of the messages are associated with geolocation metadata, which contains the coordinates of the location where the tweet is made. Unfortunately, only a few tweets contain this information, which allows us to make location-based analyses. More tweets, however, contain a string which is entered by the author. Table 5 shows the exact numbers and percentages for the tweets retrieved in July, 2011.

Metadata	Number	Percent
Geolocation	52138	0.72%
Location string	5043614	70.12%
Detected location	2539784	35.31%
Total	7193256	100.00%

Table 5: Tweets with location information

Converting a free-formed location description string to structured location information, however, is not 100% correct. This is because of a couple of reasons. First, not all users enter real locations. Some of them use that field only to enter made up texts. Some of such examples are: “Mars“, “in your mind” and “closer to my dream”. Some locations, on the other hand, can be ambiguous because of the lack of details. For example, a user may enter only their street name which can exist in more than one city. Spelling mistakes and similar errors can also cause the location resolving to be unsuccessful.

Two other types of location values are also interested. First, a person can live in more than one city and enter values similar to “Amsterdam and London” to that field. In such cases, if one of the cities is predicted, the result is considered to be correct. Second, some of the social media applications used by Twitter users attach coordinate values to the location string, instead of setting geolocation fields. This kind of locations can also be determined and converted to cities, states and countries by this component.

We have tried to determine the correctness of these predictions. Since it is not possible to check all the location values entered in all tweets, we have only focused on those tweets which are predicted to be made in one of the five

major Dutch cities. For all input strings and outputs, a human judge checked if the output matches the input; and the precision has been calculated. Recall of the operation has not been calculated, because analyzing millions of tweets manually is unpractical. Table 6 shows the types of location values entered by users, and the correctness of the operation for the cities of our interest.

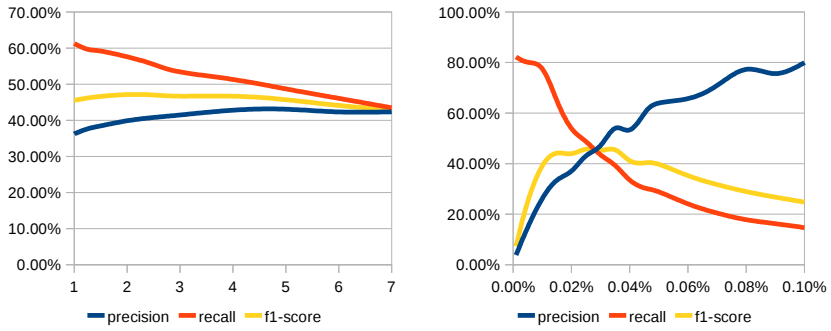
	Number of Tweets					
City	Coord.	Correct	Double	Wrong	Made up	Precision
Amsterdam	1022	11730	151	5	2	99.95%
Eindhoven	76	2070	6	9	0	99.58%
Rotterdam	397	5044	60	6	20	99.53%
The Hague	197	2797	13	0	1	99.97%
Utrecht	61	2685	23	145	0	95.02%
Total	1753	24326	253	165	23	99.29%

Table 6: Precision of detected cities

6.1.3 Term Extraction

In order to find the related terms for SEM, term extraction method explained in Section 5.3.1 has been applied. We have first started with looking for good values for m and n parameters in Formula 2 in page 24. Since the term extraction will be used for finding advertising keywords, a number of related and frequent keywords are selected as relevant. Then, the effect of these variables is observed in terms of precision, recall and F1 score. Note that, this score is computed as the harmonic mean of precision and recall as in Formula 4. The effect of m when $n = 0.025\%$ and the effect of n when $m = 5$ are shown in Figures 15a and 15b, respectively.

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$



(a) ratio parameter(m) (b) constant parameter(n)

Figure 15: Effect of term extraction thresholds in Formula 2

This method has been applied to find terms in Vodafone context. The most frequent results obtained by Formula 3, where $m = 5$ and $n = 0.005\%$ are listed in Table 13.

6.1.4 Ontology Learning

In order to find the performance of our ontology learning method, we have manually developed an ontology with 31 relationships of 5 types. For these 31 pair of words, the most frequent 40,000 patterns have been used as features and pair-pattern matrix has been constructed. Then, 10-fold cross-validation has been performed for ten times. In each iteration, rows are randomized to have different foldings. In each fold, a number of features with the most information gain are selected, and then a classification algorithm is applied. The performances of different algorithms are shown in Table 7.

Algorithm	Percent correct	Kappa
Baseline	22.58%	0.00
SMO (support vector)	37.25%	0.20
J48 decision tree	39.67%	0.27
Logistic Regression	42.33%	0.24
Naive Bayes	47.83%	0.30
Naive Bayes with kernel density estimation	50.33%	0.33
Naive Bayes Multinomial	52.58%	0.38

Table 7: Performances of classification algorithms for LRA

In order to determine the effect of selected number of features, another experiment has been made. Figure 16 shows the change of accuracy with respect to the number of patterns used. The classification algorithm used here is a Naive Bayes Multinomial classifier.

For a classification with 150 attributes, where the percentage of correct classified relationships is 58%, the confusion matrix given in Table 8 has been obtained. It is worth to note that, because of the ontology we have specified for this context, brands names (i.e. Apple) and product names (i.e. iPhone) are different types of entities. However, in real life these words can be used interchangeably and share similar language patterns. That is the most likely reason that there is some confusion between class b and class e.

The results show that, this method itself is not sufficient to learn a complete ontology automatically because the accuracy is less then desired. However, it is obviously above the majority baseline and can be helpful for domain experts who are manually developing an ontology.

If a fully automated ontology learning needs to be implemented, more features can be used other than patterns. For example, the co-occurrence of a word with other words in each category can be used as another feature for the classification problem. On the other hand, special attention must be paid to avoid using relationships with very similar semantics.

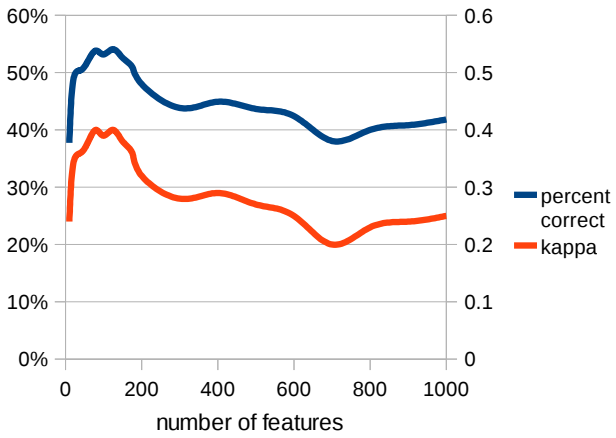


Figure 16: Ontology learning performance

	classified as					
	a	b	c	d	e	class description
real a	1	1	0	1	0	companyTOcompetitor
real b	0	3	1	0	3	companyTObrand
real c	0	0	4	2	1	companyTOcampaign
real d	0	1	0	6	0	companyTOservice
real e	0	2	0	1	4	companyTOproduct

Table 8: Confusion Matrix of LRA

6.2 Case Study: Vodafone

In order to try our analysis methods in a real-world setting, we have communicated with a major mobile phone operator, Vodafone Netherlands. We have received their web analytics data to perform some analyses.

6.2.1 Keyword Generation

We have received the paid search keywords for Vodafone Netherlands. In our analysis, it has been observed that, the keyword groups have particular patterns. Also the keywords in each group are semantically related. Some of these ad groups include the ads for following types:

- generic company: vodafone netherlands, vodaphone
- specific product: apple iphone 4, samsung galaxy tab, htc wildfire
- product family: apple iphone
- product group: smartphone, android devices, tablet pc
- contract types: prepaid, contract, sim only

- product-contract type pair: prepaid nokia c3
- specific service: cheap sms, internet, roaming

However, managing these groups with thousands of keywords manually seems to be not a straightforward task. There are some ad groups with unrelated keywords. There are also some ad groups with not enough keywords. That is why; our keyword generation method explained in Section 5.3 is considered to be useful for this setting.

Considering the keyword sets obtained from Vodafone, we also suggest an ontology with following entity and relationship types:

- company: vodafone
- product: wildfire, iphone4, smartphone, phone
- brand: apple, blackberry, htc
- service: sms, 3g, internet, coverage
- competitor: kpn, t-mobile, hi
- campaignWords: deal, offer, free, gratis, new, campaign, cheap
- productOF: iphone → apple, wildfire → htc
- isA: iphone4 → iphone → smartphone → phone

It would be also interesting and useful to allow an entity to have more than one term. This allows the use of synonyms, alternative forms and misspellings. For example, "blackberry" entity can also include its abbreviated form "bb". Similarly, "vodafone" entity can include its common misspellings such as "vodaphone" and "vodafon". Note that, it is possible to find such misspellings using our framework by computing edit distance between words.

The performances of term extraction and ontology learning regarding Vodafone data have already been given in the previous section. After term extraction and ontology development steps, we have also created an example advertising campaign structure. This is the combination of manual steps 3-4 of our campaign and keyword generation method as described in Section 5.3. The example campaign can be seen in Table 14 in page 50. We would like to emphasize the possibility of generating many keywords with a single pattern and the easiness of managing changes and updates with this method.

6.2.2 Brands and Cities

Our framework collects Twitter posts about Vodafone and its main products. Exact tracking keywords are *vodafone, iphone, ipad, htc, nokia, blackberry* and *samsung*. After retrieving these messages, location analyzer component tries to determine the location of the author, and polarity classifier component determines the sentimental polarity of the messages.

Among all the messages whose authors are determined to live in the Netherlands, the analyzer computes the share of mentions of the products for each city. Table 9 shows the results for five major cities in the Netherlands.

	iphone	ipad	black berry	sam sung	nokia	htc	total
Amsterdam	57.4%	19.3%	11.7%	5.4%	3.0%	3.3%	100%
Rotterdam	55.8%	10.2%	22.7%	4.3%	2.3%	4.8%	100%
Den Haag	61.8%	10.7%	13.7%	5.2%	3.0%	5.6%	100%
Utrecht	66.9%	8.9%	6.2%	6.2%	2.7%	9.1%	100%
Eindhoven	52.1%	14.0%	12.3%	6.3%	6.8%	8.5%	100%

Table 9: Twitter mentions of products for each city

With this table, however, it is not quite straightforward to make a direct decision. Because, the popularity of each brand is different, and the values in the table does not represent the relative popularity for the city. That is why; these values have been normalized with Equation 5. In this equation, p denotes the fraction of a particular brand in a city, p_{avg} denotes the overall share of mentions of a brand in all cities and p_{norm} is a value between -1 and 1 which shows the normalized interest in a product in a city relative to the overall interest.

$$p_{\text{norm}} = \begin{cases} \dfrac{p - p_{\text{avg}}}{1 - p_{\text{avg}}} & \text{if } p \geq p_{\text{avg}} \\ \\ \dfrac{p - p_{\text{avg}}}{p_{\text{avg}}} & \text{if } p < p_{\text{avg}} \end{cases} \tag{5}$$

Using this formula, relative interests have been calculated. Table 10 shows the results. Each cell in this table contains the p_{norm} calculated by p from the corresponding cell of Table 9.

	iphone	ipad	blackberry	samsung	nokia	htc
Amsterdam	-0.011	0.042	-0.114	0.001	-0.040	-0.286
Rotterdam	-0.039	-0.352	0.110	-0.194	-0.276	0.002
The Hague	0.090	-0.318	0.006	-0.027	-0.047	0.010
Utrecht	0.211	-0.434	-0.531	0.009	-0.125	0.047
Eindhoven	-0.102	-0.113	-0.069	0.010	0.038	0.041

Table 10: Normalized product interests for each city

Besides the shares of social media mentions, the sentiments of these messages have also been computed for each city and product. The results can be found in Table 11.

In order to verify the relevancy of these social media metrics, they have been compared with real data obtained from Vodafone. The raw web analytics data exported from SiteCatalyst implementation of Vodafone have been used to construct the following aggregated tables: sales numbers, search numbers, conversion rate. Please note that, it was not possible with Vodafone's web analytics system to get conversion rates or exact number of visits for each city.

	iphone	ipad	blackberry	samsung	nokia	htc
Amsterdam	57.2%	58.3%	47.2%	92.5%	56.2%	82.7%
Eindhoven	60.1%	89.8%	46.5%	68.2%	33.3%	80.0%
Rotterdam	65.2%	75.3%	36.7%	88.2%	55.6%	86.8%
The Hague	64.6%	79.6%	53.6%	76.9%	66.7%	78.6%
Utrecht	61.4%	76.9%	33.3%	74.1%	66.7%	80.0%

Table 11: Percentage of positive messages for each product

Therefore, it is estimated using sales and search numbers. See the corresponding tables in Appendix E.

When we compare these results intuitively, a correlation can be noticed between sales data and social media mentions. For instance, Blackberry devices are detected to be the most popular in Rotterdam. Similarly, HTC is favored in Eindhoven and Utrecht. On the other hand, the following products are detected to be unpopular in their corresponding cities: iPad–Rotterdam, Blackberry–Utrecht, Nokia–Rotterdam and HTC–Amsterdam.

Besides this intuitive observation, we have made a more formal analysis to compare these different data sources. Pearson correlation coefficient[25] has been calculated between the tables for online sales, conversion rate, number of searches, social media mentions, and sentiment. The correlation coefficient for two data samples have been calculated using Formula 6. Note that, in this formula, $\bar{X}$, s_X and r are sample mean, standard deviation and the correlation coefficient, respectively.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \tag{6}$$

The calculated correlation metrics are given in Table 12. Each cell represents the correlation between the two tables in the column and the row. For example, Mentions-Sales value has been obtained by calculating Pearson correlation between Table 10 and sales table. As it can be seen, a medium to strong positive correlation has been detected between social media mentions with online sales and conversion rate. On the other hand, search numbers and social media sentiment have not been detected to be correlated with sales and conversion.

	Sales	Conversion	Search	Mentions	Sentiment
Sales	1.000				
Conversion	0.019	1.000			
Search	-0.031	0.057	1.000		
Mentions	0.488	0.192	-0.016	1.000	
Sentiment	0.024	-0.062	-0.092	-0.038	1.000

Table 12: Correlation between different data sources

[25] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

7 Conclusions

The conclusions and discussion for this master project are given in this section. First, the main contributions of the project are listed. Second, its possible implications for the industry are analyzed. Then, the limitations and drawbacks of the current work are discussed. Finally, some possible improvements and extensions are listed which can be possible directions for future work.

7.1 Main Contributions

A concise overview of social media measurement, web analytics and search engine marketing fields have been provided by describing related concepts and state-of-the-art. The gap between social media and the two other fields has been pointed out and suggestions have been made for bridging this gap.

In order to measure social media data and align it with the web analytics data from other resources, a framework has been suggested and implemented. The resulting product is capable of retrieving public social media entries from Twitter, blogs and news sites. Data is preprocessed and associated with metadata such as sentimental polarity and location of the author. It is possible to filter, segment and aggregate this data in a similar way to web analytics data. Combining social media data with web analytics data enables new types of analyses which are not possible with web analytics data only. Combined results and reports can be considered by marketing executives in their decision making process.

Two types of data analysis operations have been performed on social media data using the framework. The results of these analyses are intended to improve search engine marketing by making use of social media data. First, a new method has been suggested to create and manage marketing campaigns. According to this method, social media messages are used as a text corpus to extract relevant terms and to construct an ontology in the domain of a company. This domain knowledge enables marketers to create generic ad groups and keyword patterns to expand their advertising keyword sets. By having a complete set of related keywords and using cheaper ones, they can reduce marketing costs. As a second type of analysis, a hypothesis has been introduced, which suggests that the interest of customers in different products may change locally. This hypothesis has been tested by analyzing social media data for several products in different geographical regions and comparing it with web analytics and sales data for these location-product pairs.

As a result of the analysis made, social media mentions information has been shown to be correlated with and useful for search engine marketing. This information and the results of social media measurement can be used to make easier marketing decisions.

The developed framework can be integrated with current web analytics products to align social media data with web analytics data. In order to verify this claim, integrations has been made with two major web analytics solutions, namely Google Analytics and Adobe SiteCatalyst.

An existing technique, LRA has been adapted and applied for the first time in ontology learning domain. According to this method, a text corpus is used to extract natural language phrases and to identify latent relationships between

words in those phrases. The types of relationships between pairs of words can be found using this method and an incomplete ontology definition can be extended.

7.2 Implications

This piece of work has three types of implications: scientific, industrial and engineering. All of these types are explained below:

7.2.1 Scientific Implications

Scientific implications are academic contributions to the literature and possible future directions of this work. The first contribution is the use of LRA method for ontology learning purposes. This method can be used by other studies and adapted to different domains. Also social media data has been used as a corpus for the first time for this kind of problems, to the best of our knowledge. This idea can also be picked up by other researchers.

On the other hand, campaign optimization problem has been handled differently compared to the previous work. Keyword generation problem for campaigns used to be done either on single-word level, or by making use of search engine logs. We have suggested a new method on multi-word level and by semantically analyzing a data corpus. Moreover, location parameter of search engine marketing campaigns has been considered for the first time by an academic study, according to our literature survey.

7.2.2 Industrial Implications

Industrial implications consist of the methods developed to create and improve search engine marketing campaigns. Our semi-automated method which makes use of domain knowledge, makes it easier to manage big campaigns with many keywords. This approach can be taken over by companies in their campaign management processes.

Another business-related implication of this project is the use of social media data for marketing purposes. Our local product interest analysis and relevant search term extraction methods can be used by marketing experts. The idea of focusing different product campaigns on specific regions can also be helpful to improve marketing campaigns even without a social media analysis.

7.2.3 Engineering Implications

The design principles and implementation methods used in this project make engineering implications for the developers of similar products. For instance, our SM and WA data alignment method, the architecture and component structure of the framework, and integration methods can be used by those who develop web analytics or social media measurement solutions.

Moreover, some of the components use interesting techniques to perform their tasks. For example, news and blog data is retrieved by Google Alerts feeds. On the other hand, free-formed locations are converted into structured locations by making use of geolocation API's. These ideas can also be useful for some software engineering companies.

7.3 Limitations

The limitations of this project have been classified into three broad categories. The first category includes those originated from computational resource limitations. Second and third category contains methodological and practical limitations, respectively.

7.3.1 Computational Limitations

The framework has been designed to work on a single node, which limits the scalability on a number of levels. First, the amount of data which can be retrieved is limited. This is a result of both the scarcity of computing resources and the artificial rate-limits applied by third parties whose API's are being used. These API's include Twitter search and streaming API's and geolocation API's of Google and Yahoo. For this reason, a limited number of messages can be retrieved and preprocessed by a single node.

Second scalability limitation results from the size of the database. As more and more data is retrieved, it becomes more difficult to store all the data on a single node. Besides storage restrictions, querying and analyzing this data becomes a non-trivial task.

These scalability limitations can be overcome with some improvements. First, data retrieval and preprocessing problems can be solved by distributing these tasks on several nodes. For example, a load balancer component can be developed to monitor and distribute the tasks. This way, different social media users and tracking keywords can be handled by different nodes of the system. Second, a scalable non-relational data store can be used instead of an SQL database. After modifying the algorithms to work with a scalable approach, by using map-reduce operations for instance, data query and analysis problems can be solved.

7.3.2 Methodological Limitations

A methodological limitation of the framework is the completeness and integrity of the data. Web analytics systems do not usually have complete data, because they are affected by user settings. For example, users can share different IP's and browsers; a computer can be shared by many users; and cookies can be blocked or removed. Because of these possibilities, determining which requests belong to the same user is not straightforward. Moreover, most web analytics tools only work with browser which has JavaScript enabled. Therefore, web analytics data may be incomplete or inaccurate.

Similar to the web analytics data, the retrieved social media data may not be complete. First, because of privacy settings of users, only public messages can be retrieved. Second, keyword-based filtering may not cover all relevant messages and recall may be lower than 100%. Third, spam messages published by some users may affect number of mentions, sentiment, and similar results.

7.3.3 Practical Limitations

Current solution may not be cost-efficient to implement for small companies, because they have to setup their own system. It is less effective in terms of server and labor costs. For a commercial application of this framework, SaaS model

can be used. A centralized system can serve for multiple clients simultaneously in order to reduce overall use of resources.

7.4 Possible Improvements and Extensions

Currently, sentiment analysis is done rather superficially. Only unigrams are selected as features and Naive Bayes classifier is being used. This method can be improved by using more advanced and hybrid classifiers, making better training sets and using other indicators as features. The coverage of languages may also be expanded. Currently, sentiment analysis is only supported in two languages: English and Dutch. Classifier models can also be built for other languages in order to cover more users and the contents they share.

Regarding ontology learning, the results show that the current method is not accurate enough to be used without human intervention. It may be possible to improve these results by considering more techniques other than LRA. For example, new features can be added to pair-pattern matrix to consider different kinds of information such as co-occurrence with the words of different entity types.

In our implementation, the coverage of data sources is limited. Only Twitter "tweets", blog posts and news articles are followed and analyzed. By implementing special data retrievers for missing social media platforms, the data can be more complete. Some of the useful platforms which can be followed include Facebook, LinkedIn, Google Buzz and some local social networks such as Hyves.

As its current situation, this project cannot be used as an end-to-end solution to create SEM campaigns. To improve this, it can be integrated to marketing decision support engines such as O2mc. Currently, those tools consider data from several resources in order to optimize online marketing campaigns. Social media statistics and sentiments can be an important contribution to these resources.

References

[1] V. Abhishek and K. Hosanagar. Keyword generation for search engine advertising using semantic similarity between terms. In *Proceedings of the ninth international conference on Electronic commerce*, pages 89–94. ACM, 2007.

[2] M. Baroni and S. Bisi. Using cooccurrence statistics and the web to discover synonyms in a technical language. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume 5, pages 1725–1728. Citeseer, 2004.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

[4] P. Buitelaar, D. Olejnik, and M. Sintek. A protégé plug-in for ontology extraction from text based on linguistic analysis. *The Semantic Web: Research and Applications*, pages 31–44, 2004.

[5] P. Buitelaar, P. Cimiano, and B. Magnini. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12, 2005.

[6] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 659–664. Citeseer, 2005.

[7] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4): 407–422, 2005. ISSN 0893-6080.

[8] X. Ding, B. Liu, and L. Zhang. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1125–1134. ACM, 2009.

[9] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords, 2005.

[10] J. Feldman, S. Muthukrishnan, M. Pal, and C. Stein. Budget optimization in search-based advertising auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 40–49. ACM, 2007.

[11] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *Proceeding of the 17th international conference on World Wide Web*, pages 61–70. ACM, 2008.

[12] M. Ganapathibhotla and B. Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 241–248. Association for Computational Linguistics, 2008.

[13] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009.

[14] T.R. Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[15] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.

[16] N. Jindal and B. Liu. Mining comparative sentences and relations. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1331. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[17] J. Krauss, S. Nann, D. Simon, K. Fischbach, and P.A. Gloor. Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS European Conference on Information Systems*, 2008.

[18] S. Muthukrishnan, M. Pal, and Z. Svitkina. Stochastic models for budget optimization in search-based advertising. *Internet and Network Economics*, pages 131–142, 2007.

[19] P. Nakov and M.A. Hearst. Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL*, volume 8. Citeseer, 2008.

[20] R. Narayanan, B. Liu, and A. Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 180–189. Association for Computational Linguistics, 2009.

[21] B. OConnor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

[22] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC 2010*, 2010.

[23] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[24] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[25] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pacific Symposium on Biocomputing 2002: Kauai, Hawaii, 3-7 January 2002*, page 362. World Scientific Pub Co Inc, 2001.

[26] P.v.d. Reijden and O. Koppius. The value of online product buzz in sales forecasting. In *Proceedings of ICIS*, 2010. URL `http://aisel.aisnet.org/icis2010_submissions/171`.

[27] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter. Edgar: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 517. NIH Public Access, 2000.

[28] P. Rusmevichientong and D.P. Williamson. An adaptive algorithm for selecting profitable keywords for search-based advertising services. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 260–269. ACM, 2006.

[29] P. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth european conference on machine learning (ecml-2001)*, 2001.

[30] P.D. Turney. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1136–1141. Morgan Kaufmann Publishers Inc., 2005.

[31] P.D. Turney and P. Pantel. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[32] Peter Turney. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL `http://www.aclweb.org/anthology/C08-1114`.

[33] S. Vintar, L. Todorovski, D. Sonntag, and P. Buitelaar. Evaluating context features for medical relation mining. In *ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics*. Citeseer, 2003.

[34] W. Yih, J. Goodman, and V.R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th international conference on World Wide Web*, pages 213–222. ACM, 2006.

[35] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 213–222, New York, NY, USA, 2006. ACM.

A Implementation Details

The components and data flow of the system are presented in Figures 11 and 12 in page 21. The details of each component can be found below:

A.1 Data Retriever

Data retriever component connects to the external entities in order to retrieve the necessary data required by the system to perform its job. Most of the data used in this project is retrieved by this component.

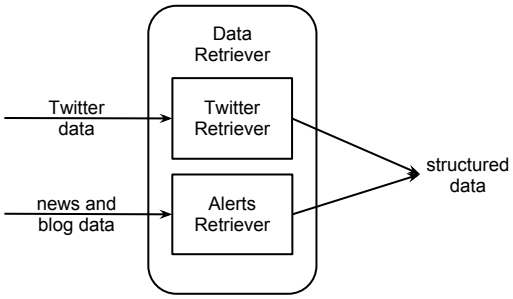


Figure 17: Data retriever component

A.1.1 Twitter Data Retriever

Related status messages of Twitter users are retrieved by this component. There are two different application programming interface (API)'s which can be used to do this task. Namely, Twitter Search API and Twitter Streaming API. Both techniques have some advantages and disadvantages. Search API returns historical data back to at most one week. On the other hand, Streaming API provides real-time results. Streaming API gives more metadata about tweets, for example, location and timezone of the user, number of followers and statuses of the user, etc. It is also possible to determine annotations like retweets, mentions and replies more reliable than Search API.

On the other hand, Search API can be used periodically to collect data similarly to the Streaming API. Some metadata is missing, but it is possible to get historical data. Therefore, when a query is made for the first time, Search API can be used. Moreover, since Streaming API requires an open connection to Twitter servers, which may stop because of network problems or other issues, Search API can be used as a backup plan to retrieve missing data. Search API also provides language identification of the tweets which cannot be accessed using Streaming API.

In our implementation, Search API and Streaming API are both used in order to capture all the available information reliably. A single instance Java application is connected to the Streaming API to retrieve real-time data. Since the Twitter connection may be broken, a new instance of the application is created every hour, which internally terminates the previous instance. Moreover, another module of the application connects to the Search API on a daily basis

in order to capture any missing data and also historical data when a new query will be followed. The data retrieved from the two parts of this component is saved into the *Tweet* table of the database.

A.1.2 Alerts Listener

Google Alerts is a search-related tool provided by Google search engine. Once a query is defined with a Google account, this tool provides updates to a feed whenever Google indexes a new webpage which matches the query given. Different kind of websites can be selected and followed such as news and blog sites. Since those websites may contain related information to the system, a component is developed to retrieve data from Google Alerts.

Although it is possible to poll an Atom feed to get recent items, another method simplifies the process. This is called PubSubHubBub. In this scenario, a client (this component) subscribes to a feed publisher (Google Alerts) via a third-party tool called a Hub. Hub follows the publisher and whenever a new item is published, it distributes the data to all its subscribers. Therefore, subscribers do not have to poll the publisher, which is a waste of resources for both parties.

This component provides methods to register a Google Alerts feed via PubSubHubBub protocol. It also has a web server which receives, parses and saves the updates into the database when a new item is served by the Hub via HTTP-Post protocol. The data retrieved by this component is stored into the *Message* table.

A.2 Data Preprocessor

The preprocessing of the retrieved data is done by this module. It basically associates some meta information such as location and polarity to the messages. There are two submodules, both of which are responsible of different preprocessing tasks.

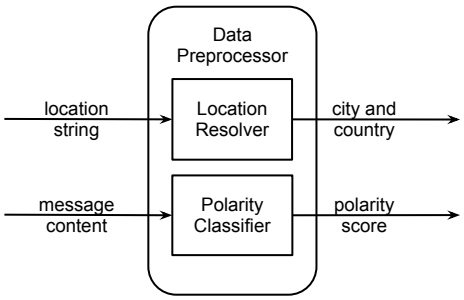


Figure 18: Data preprocessor component

A.2.1 Location Resolver

The tweets retrieved from Twitter API's do not have location information in terms of countries and cities. There are just two types of information which can

be used. First, some of the messages have geo-tags, which contain the latitude and longitude of the users. Second, users are able to enter a free-form text into a field called "location". Although some users enter irrelevant texts in this field, it is mostly useful to determine the exact location of a user.

In this project, geolocation API's of Google and Yahoo are used in order to convert locations into a structured form which keeps the country, state and the city. If a coordinate is given by means of a geotag, it is possible to find the exact city, state and country. Otherwise, location field is given as an address to the API's and if a match is found, the result is used as the actual location information.

A.2.2 Polarity Classifier

WEKA data mining library is used to apply classification algorithms. NaiveBayesMultinomial classifier has been chosen because of its simplicity and promising results. The classifier models have been built using the training sets for English and Dutch languages. Then feature extractor (filter) and the model have been saved to use in future classification tasks. When new messages are need to be classified, classifier model is loaded and the trained Naive Bayes classifier is used to determine the polarity of the messages.

A.3 Data Selector

The raw data and associated metadata are stored in the database by other modules. Parts of this data is also need to be loaded in order to aggregate and display the results and perform analysis. Since one of the aims of the projects is to align social media data with web analytics data, this component is developed in such a way that, it provides an object oriented abstraction of the database similar to web analytics API's in the market. In the examples below, the design of the interface of this component can be seen in different scenarios:

Figure 19: Data selector component

A.3.1 Filtered Messages with Metadata

```
DataSelector ds = new DataSelector();
ds.addFilter(Filter.COUNTRY, "NL");
ds.setMetadataNeeded(Metadata.SENTIMENT);
ds.getMessages();
```

A.3.2 Time Series of a Metric

```
DataSelector ds = new DataSelector();
ds.addFilter(Filter.COUNTRY, "NL");
```

```
ds.addFilter(Filter.SINCE, since);
ds.addFilter(Filter.UNTIL, until);
ds.addMetric(Metric.NUM_MESSAGES);
ds.addMetric(Metric.NUM_POSITIVE);
ds.getTimeSeries();
```

A.3.3 Metrics and Dimensions

```
DataSelector ds = new DataSelector();
ds.addMetric(Metric.NUM_MESSAGES);
ds.addMetric(Metric.NUM_USERS);
ds.addFilter(Filter.SINCE, since);
ds.addFilter(Filter.COUNTRY, "NL");
ds.addFilter(Filter.CONTAINS, "iphone");
ds.addDimension(Filter.STATE);
ds.addDimension(Filter.CITY);
ds.sortBy(Metric.NUM_MESSAGES, false);
ds.getMetricsAndDimensions();
```

A.4 Data Analyzer

Data analyzer is the key component of the system. It gets data compiled by the other components and performs the analysis tasks defined in the project goals. All of the sub-components seen in Figure 20, implements the corresponding methods explained in Section 5.

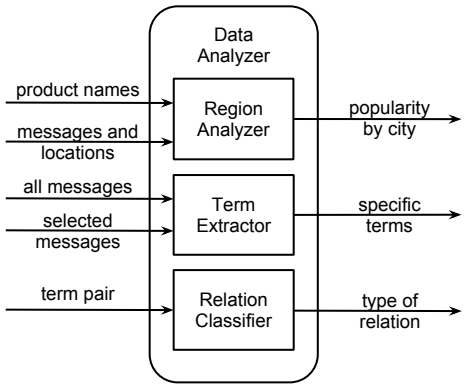


Figure 20: Data analyzer component

A.5 Integration with Web Analytics Tools

In order to explore the integration possibilities between the social media framework and web analytics applications, some possibilities have been explored. Two integrations possibilities have been identified. First, the social media data can be exported to the web analytics tool to display social media data in web analyt-

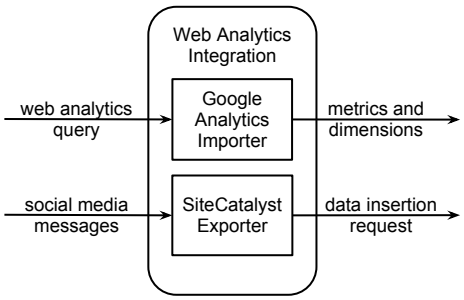


Figure 21: Web analytics integration component

ics reports. Second, web analytics data can be imported from the web analytics tools to display it along with social media statistics.

The first type of integration has been made between the project and Adobe SiteCatalyst. With their Java AppMeasurement library, it is possible to upload data to web analytics data servers. For each social media message, the following information is sent: source, author, polarity, country, state, city and time. As a result, aggregated social media data can be seen in SiteCatalyst reports, including number of mentions timeline, shares of social media sources, percentage of positive and negative messages and distribution of cities.

This integration has some limitations. First, the current version of SiteCatalyst does not support advanced segmentation and breaking the data down to different dimensions. Therefore, limited functionality is available. More advanced data analysis tasks should be performed within the framework. Second, because of access restrictions, social media data is not combined and shown together with real-world data of a company. Third, it is not possible to track broad keyword with this integration enabled, because Adobe charges its customers based on the number of website hits, and with this method, every social media message is processed as a webpage hit. Therefore, following social media within SiteCatalyst can bring a significant additional cost. In order to overcome this problem, only aggregated data can be uploaded to SiteCatalyst instead of individual messages. But this method restricts the types of possible analyses even further.

Second type of integration has been made between our framework and Google Analytics. It is possible to retrieve web analytics data from a Google Analytics account using its Data Export API[26]. It provides sufficient flexibility for our analyses. It is possible to use this API to apply similar dimensions and metrics to web analytics data to social media data. For example, we can show two time series on a single chart comparing the webpage views and social media mentions for a particular product and city.

[26] `http://code.google.com/apis/analytics/docs/gdata/home.html`

B Data Model

The database model of the framework is depicted in Figure 22.

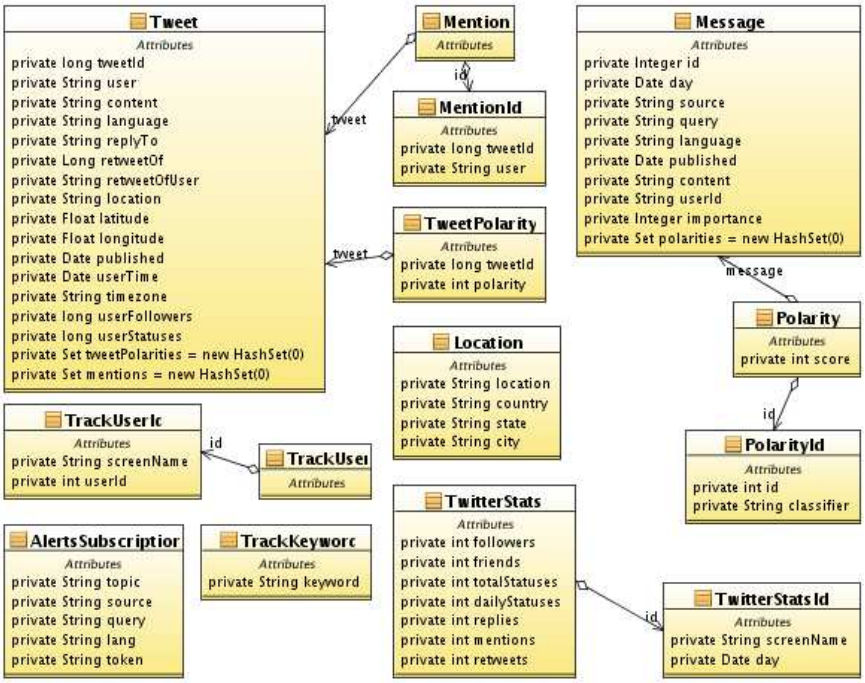


Figure 22: Data Model

C Keyword Set Expansion Techniques

This section describes several methods to identify word relationships and extend keyword sets. The details can be found below.

- **AdWords:**. Google provides its own tool to suggest keyword ideas to AdWords users. Given an initial set of keywords, it provides similar ones using its own data. But since a lot of users use this tool, the generated keywords are not unique and therefore average CPC values are relatively high. A similar keyword suggestion method using search engine click log is suggested by Yih et al [35].
- **WordNet:** WordNet is a lexical database which contains English words and their semantic relations. It is possible to find synonym, antonym, hypernym words, sister terms, etc. However, it only contains dictionary words and only applicable to English.
- **Co-occurrence:** Extracting patterns and counting co-occurrence to calculate a similarity metric (like cosine similarity) can be useful. But a method solely based on co-occurrence may not be useful enough. Because, exact types of relationships is very hard to identify. Abhishek [1] used this method with a corpus of websites in a particular domain (healthcare).
- **Pattern:** It is possible to manually generate a list of some natural language patterns and try to capture them in a corpus. For example, "X vs Y", "X is better than Y" patterns can be used to find competitors or substitutes and "X of Y", "X's Y" can be used to extract features/parts of a word. However, this needs manual labor and cannot be generalized for all languages.
- **AltaVista:** Turney suggests a method to use AltaVista search engine to find synonyms of a given word [29]. But the scope is limited to semantically similar words based on co-occurrence.
- **Lucent:** Apache Lucent can be used to get common misspellings with its proximity search function. It is also possible to use its indexing facility for latent semantic analysis (LSA) and LRA applications.
- **LSA:** This method uses term-document occurrence matrix and reduces its dimensions using singular value decomposition (SVD) technique. Based on this matrix, similar terms can be identified.
- **LRA:** Turney suggest another method to automatically capture natural language structures from a corpus in order to identify several types of relationships between words [32].

D Ontology Results

word	freq. in vodafone context	freq. in general context	word	freq. in vodafone context	freq. in general context
vodafone	8.498%	0.000%	**sony**	0.111%	0.001%
phone	0.571%	0.108%	update	0.111%	0.013%
mobile	0.524%	0.005%	**tab**	0.108%	0.001%
iphone	0.434%	0.044%	**kpn**	0.107%	0.000%
htc	0.410%	0.001%	sfr	0.107%	0.000%
3g	0.409%	0.001%	**ericsson**	0.106%	0.000%
uk	0.405%	0.057%	**deal**	0.105%	0.019%
free	0.389%	0.051%	customers	0.105%	0.002%
internet	0.372%	0.022%	bill	0.103%	0.005%
service	0.327%	0.013%	**net**	0.101%	0.006%
blackberry	0.270%	0.012%	essar	0.097%	0.000%
network	0.260%	0.007%	moet	0.096%	0.000%
sensation	0.218%	0.000%	**prepaid**	0.096%	0.000%
galaxy	0.208%	0.001%	**phones**	0.096%	0.008%
sms	0.201%	0.001%	**services**	0.094%	0.002%
data	0.196%	0.004%	**calls**	0.093%	0.010%
customer	0.190%	0.004%	airtel	0.093%	0.000%
store	0.189%	0.017%	**desire**	0.092%	0.001%
t-mobile	0.185%	0.000%	launch	0.091%	0.004%
samsung	0.167%	0.001%	**price**	0.084%	0.010%
sim	0.164%	0.004%	con	0.084%	0.008%
xperia	0.161%	0.000%	dus	0.083%	0.001%
android	0.160%	0.004%	turkcell	0.082%	0.000%
app	0.156%	0.017%	**smart**	0.082%	0.009%
pay	0.154%	0.021%	doet	0.080%	0.000%
bb	0.152%	0.023%	australia	0.079%	0.010%
contract	0.150%	0.002%	**apple**	0.078%	0.012%
number	0.145%	0.023%	zijn	0.078%	0.000%
white	0.143%	0.025%	recharge	0.078%	0.000%
o2	0.138%	0.002%	pre-order	0.077%	0.001%
india	0.130%	0.003%	**deals**	0.074%	0.002%
nokia	0.126%	0.001%	ad	0.073%	0.006%
per	0.125%	0.004%	**coverage**	0.072%	0.001%
orange	0.122%	0.009%	**nl**	0.072%	0.001%
vivendi	0.121%	0.000%	**winkel**	0.071%	0.000%
stake	0.118%	0.000%	**smartphone**	0.071%	0.000%
buzz	0.116%	0.002%	courtesy	0.071%	0.002%
available	0.113%	0.010%	**offer**	0.070%	0.006%
signal	0.111%	0.003%	euro	0.069%	0.001%

Table 13: Extracted terms from social media for "Vodafone"

group	ad text	keyword patterns	generated keywords
specific product	get the newest iphone for the best price	(apple), phone4 (apple), iphone4, [campaignWord] (apple), iphone4, [contractType]	iphone4, iphone 4, apple iphone4 iphone 4 deals, cheap iphone4 prepaid iphone4, iphone4 contract
product group	best smartphones with google android OS	groupX groupX, [campaignWord]	android, htc desire, google nexus android deals, best android, cheap htc wildfire
competitor	vodafone has the best prices in the market	[competitor] [competitor], [product]	kpn, t-mobile kpn iphone, t-mobile htc wildfire
company specific	the best mobile operator in the Netherlands	vodafone	vodafone, vodaphone, vodafone nl
service specific	surf on the internet for cheap with vodafone 3G	3g, ¡campaignWord¿ vodafone, 3g vodafone, 3g, [price]	cheap 3g, best 3g, 3g deals 3g vodafone, vodafon 3-g vodafone 3g tariff, 3g costs vodafone
generic services	visit our website for information about vodafone services	vodafone, [service]	vodafone 3g, vodafone roaming

Table 14: Sample ad campaign with patterns for Vodafone

E Vodafone Data (Confidential)

The confidential data in this section is not available in the public version of this thesis.