Sentiment Analysis on Yelp social network

Flora Amato, Giovanni Cozzolino, Antonino Mazzeo and Antonio Pizzata
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI
University of Naples "Federico II", ITALY
Email: {flora.amato, giovanni.cozzolino, mazzeo, antonio.pizzata}@unina.it

Abstract

Social networks analysis is an emerging trend among scholars and researchers in the last years. A great number of companies are interested in social networks data mining. Data gathered from Facebook, Twitter or other social networks result to be very attractive in many application fields, like economics analysis, sentiment analysis, and politics analysis and so on. In our work, we focused on the analysis of the content of textual information obtained from the social media. Our investigation is finalized to extract hot topics in social network. We considered, as case study, reviews obtained from the social network Yelp.

1 Introduction

Nowadays, social network information are precious for many application fields. For example information gathered from social network are used by so-called "social media marketing" to organize an advertising campaign: companies can reach a bigger number of stakeholders and create a more modern concept of advertising, quicker, immediate and sometimes matched with catchy hashtags.

Important information that analyst are interested in, regarding how people feel about an event occurred in their lives or around them. It has been proved [21] that the use of emoticon or hashtags can identify the sentiments people feel: the use of these forms of communications are the easiest way to share them on the net. This led researchers and scholars to extend the study of sentiment analysis[14, 15] to the digital world, mining information through forums, blogs, social media, in order to understand how the people react to and event (i.e. a new law, terrorist attack and so on).

In literature, there are many tools and projects designed to perform different kind of analysis[4].

In our work, we focused on the analysis of the content of textual information obtained from the social media: through our investigation, it is possible to extract hot topics in social network. We considered, as case study, reviews obtained from the social network Yelp.

1.1 Text Mining Procedures

Effective and efficient access to domain relevant information requires the ability to automatic process and organize the information especially if these are contained in huge repositories of data [7, 6]. The most used approaches in Big Data processing are based on the graph algorithms, parallel and distributed architecture.Some Big Data infrastructures deal with Apache Hadoop [26] software for data-intensive distributed applications, based in the MapReduce programming model and a Distributed File System (Hadoop). MapReduce job splits the input dataset into independent subsets that are dealt with map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to compute the result of the job. Some open source tools for Big Graph mining are proposed, as Pegasus, a big graph mining system built on top of MapReduce.

It allows to find patterns and anomalies in massive real-world graphs. Another Big Data Mining initiative is Apache a scalable machine learning and data mining open source software based mainly on Hadoop and a collection of hardware, software and design patterns for managing very fast large-scale data at very low cost and using BIDMat an interactive matrix library that integrates CPU and GPU acceleration.

For what concerns the text analysis, Morphosemantic approaches similar to the one proposed here have been already proposed for many languages and applied to the medical domain. Works that deserve to be mentioned are Pratt on the identification and on the transformation of terminal morphemes in the English medical dictionary; Wolff on the classification of the medical lexicon based on formative elements of Latin and Greek origin; Pacak et al. on the diseases words ending in -itis; Norton e Pacak on the surgical operation words ending in -ectomy or -stomy; Dujols et al.on the suffix -osis. Between the nineties and the 2000, many studies have been published on the automatic popula-

DOI reference number: 10.18293/DMSVLSS2017-004

tion of thesauri, we recollect among others Lovis et al.,that derives the meaning of the words from the morphemes that compose them; Lovis et al. that identifies ICD codes in diagnoses written in different languages; Hahn et al.that segments the subwords in order to recognise and extract medical documents; and Grabar e Zweigenbaumthat uses machine learning methods on the morphological data of the thesaurus SNOMED (French, Russian, English). Several works focused the problem to the definition and the implementation of a comprehensive architecture for information structuring, while the work is dedicated to resolve the issue of ensuring semantic interoperability of different entities by mapping the content of different corpora on a set of shared concepts.

For what concerns the decision support system in literature they are usually categorized in two typologies, Knowledge-based and non-Knowledge-based [13], [25]. The firsts are accurately described in [22]. AAPHelp, created in 1972, was an early attempt to implement automated reasoning under uncertainty. Other systems are Asbru, EON and PRODIGY [32]; PROforma, SAGE [31]; and the Clinical Reminder System [20]. The last one is based on the [29] Evidence-based medicine and provides evidence-based clinical guidelines. A more detailed and systematic overview on many other CDSS is described in [19]. Recently, many studies focused on medical information extraction from structured or unstructured texts. Fette [18] presents a IE systems that integrates medical unstructured information into a clinical data warehouse to transform into a structured format the information inserted by physicians in a clinical information system. Rink [28] proposes a method for the automatic extraction of medical concepts and relations from electronic medical reports. Medical concepts are extracted with supervised machine learning algorithms. Several knowledge sources are used for feature extraction: a semantic role labeller, a POS tagger, a phrase chunk parser, WordNet, Wikipedia and the General Inquirer lexicon. Doan [17] introduced an automated system to extract medications and related information from discharge summaries. The researchers developed an integrated system adapting some existing NLP tools. In order to proper model data, several approach have been proposed [23, 24, 5]. Moreover, to efficiently process huge amount of data, several approach regarding hardware implementation of data processing tools are developed[2, 3]. In particular, in[11] and [10] authors proposed an hardware implementation of a Decision Tree based multi-classification system traffic analyzer . The system is able to deal with a huge amount of data and tight constraints, such as power consumption and hardware resources. In[9] a traffic analysis hardware accelerator, based on the Decision Tree model, is presented through an infrastructure which collects data from mobile devices and provide them update versions of the analyzer by exploit new traffic information[8]. Moreover, in the field of data protection, in [12] authors proposed a secure infrastructure to protect intellectual property installed on the FPGA by means of partial dynamic configuration.

2 A Social Network Analysis Methodology

The predominant approach to analyze social network is the graph theory, even though it is largely debated. This theory derives from the studies of Euler and provides us a way for studying Networks of any kind. In social networks the single user or groups of users are represented as a point and their relation are represented as lines. The data obtained from these graphs are then recorded in matrix form, in this way we can study directly the data without drawing the graph, that helps a lot when we are facing large social network data sets. To the lines connecting points in the diagram we can assign a direction in order to determine which point influence the other and to that influence we can also assign a value to represent the strength of that relation.

2.1 Text Analysis Tools

2.1.1 TaLTac

TaLTaC (Trattamento automatico Lessicale e Testuale per lanalisi del Contenuto di un Corpus, Lexical and Textual automatic processing for analyzing the Content of a Corpus)[30] is a software able to perform on documents and data written in natural language operations like: Text Analysis, Text Mining and Corpus Analysis. It has been developed in Italy from the conjunction of the University of Salerno and University La Sapienza of Rome. The first task to do in our environment is creating a Work Session, which is the file that is going to contain all our information, then we build the Corpus, our main object, which will be analyzed with various instruments and operations. The Corpus is then divided in two parts: fragment and section. The first is identified by four asterisks (****) a name and some variables, if needed, identified themselves with an asterisk, a name and a value. The latter is defined with four plus signs (++++) and a name. Every Corpus can contain more fragments and sections. One of the main functions of TaLTaC is the extraction of significant information from the Corpus (Text Mining), such task uses endogenous and exogenous resources: the former is composed of the number of fragments and the categorical variables which can be associated to the text in order to identify fraction of the corpus logically related. Thanks to this kind of resource TaLTaC is able to perform a Specificity Analysis. The exogenous resources are lists which contain the frequency of a term or lexical unit, thanks to these lists the software is able to identify peculiar language of the text.

2.1.2 Gate

GATE (General Architecture for Text Engineering) [27] is a open source free software which excels at Text Analysis. The first version was written in the mid-1990s and it has reached currently version 8. The main resource of this program is ANNIE (A Nearly New Information Extraction System), which provides a lot of information extraction techniques such as the English Tokeniser which splits the text into annotations of type Token or the POS Tagger that assigns to every token an annotation that describes its characteristic (i.e NNP for Proper Noun in singular form) and more. The Gate software is a family made up of:

- an IDE, Gate Developer
- a web app, Gate Teamware
- a framework, Gate Embedded

2.2 The Methodology

In order to achieve a fine analysis of our data we applied different types of operations.

2.2.1 Text Pre-treatment

This is the first process to be applied on text, in order to obtain a clear analysis in the successive processes. It is made up of the following phases. **Normalization:** This phase allows to remove any data duplication and it normalizes the writings of names, acronyms and other entities. In order to achieve these goals we have to execute various tasks:

- Change apostrophes into stresses (for the words that is needed), in order to determinate the right word;
- Label words/sequence of words so that they can assume the right meaning and are not mistook with others expressions (i.e. a name can be mistaken for a noun, Rose or rose);
- Change of capital letters into lowercase for the words that are not labeled, if one is labeled we need to analyze what the label says: if it is a name (of person, of a city or a general proper name) then it will keep the capital letter, in other cases it will not.

Correcting Spelling Errors: This phase consists of comparing a misspelled word with the system dictionary in order to correct and analyze it in the right way.

2.2.2 Lexical Analysis

Lexical analysis analyzes the segments of the Corpus (a segment is a sequence of graphic forms separated by a strong divider). A segment can be easily be defined by choosing a set of dividers (i.e. punctuation such as .,;) and then separating the sentences between these elements. Once we obtained our segments we can estimate various analysis parameter such as the IS index: this measures the level of absorption of the segment regard the single elements which it is made of. Other important operations in this part of the analysis are the Tagging, which links to every word a description of the grammatical or semantic characteristics[16, 1], and the Lexation which identifies the sequence of words defined during the pretreatment as one unique entity. Last we define the Corpus key words by studying the repetition rates , the ones that have a noticeable standard deviation (considering only the integers) can be assumed being meaningful.

2.2.3 Textual Analysis

The first step in this analysis is the study of the Concordances in which we can examine the context where every word or segment we choose is. Then we calculate the TF-IDF rate which sorts the researchs results according to the frequency and distribution of the search keyword in the documents provided. The TF-IDF is equal tf-idf $= tf \cdot \log \frac{N}{n}$ where tf is number of occurrences of an element, and the remaining part is the logarithm of the ratio between the number of documents building the Corpus (N) and the number of documents which present that element (n). Another import part of Textual Analysis is the co-occurrences identification, where with co-occurrences we identify those couple of near elements that repeat in the text. This identification is useful to define the primary concepts contained in the Corpus.

3 Experimental Campaign

Our experimentation aims to analyze online social networks data sets in order to derive as much information as possible. The data set analyzed comes from the social network Yelp, founded in 2004, which publishes crowdsourced reviews about local businesses[33] and it is made up of 50 tuples structured this way:

- anonymized user name
- anonymized reviewed place
- date of the review
- review

3.1 TaLTac Analysis

Prior starting the analysis we pass through the Text Pretreatment, and so after the parsing of the Corpus we normalize (Figure 1) it and compute the sub-occurrences.

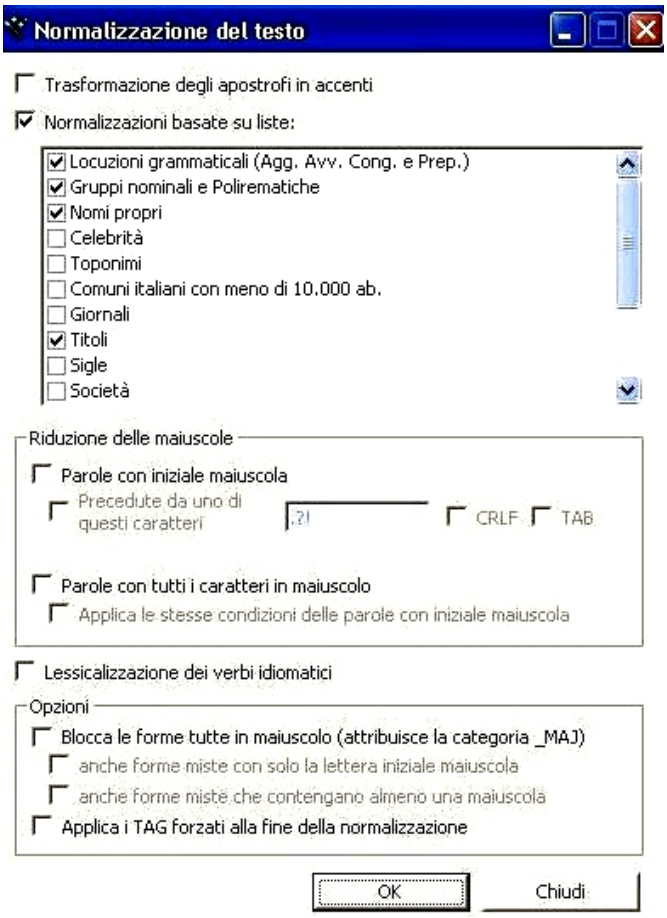
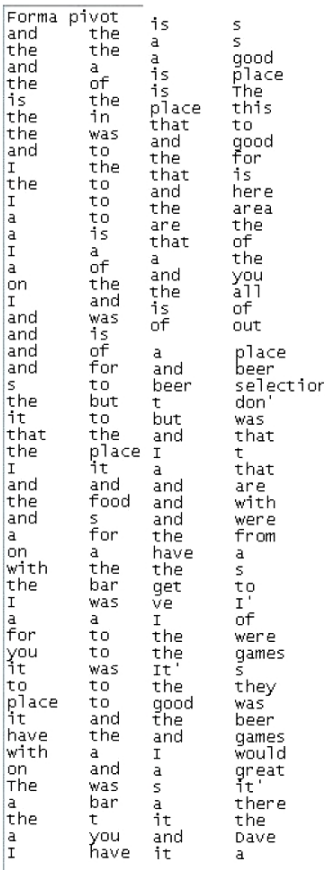


Figure 1. Text Normalization window



Forma pivot			
and	the	is	s
the	the	a	s
and	a	a	good
the	of	is	place
is	the	is	The
the	in	place	this
the	was	that	to
and	to	and	good
I	the	the	for
the	to	that	is
I	to	and	here
a	to	the	area
a	is	are	the
I	a	that	of
a	of	a	the
on	the	and	you
I	and	the	all
and	was	is	of
and	is	of	out
and	of	a	place
and	for	and	beer
s	to	beer	selectior
the	but	t	don'
it	to	but	was
that	the	and	that
the	place	I	t
I	it	a	that
and	and	and	are
the	food	and	with
and	s	and	were
a	for	the	from
on	a	have	a
with	the	the	s
the	bar	get	to
I	was	ve	I'
a	a	I	of
for	to	the	were
you	to	the	games
it	was	It'	s
to	to	the	they
place	to	good	was
it	and	the	beer
have	the	and	games
with	a	I	would
on	and	a	great
The	was	s	it'
a	bar	a	there
the	t	it	the
a	you	and	Dave
I	have	it	a

Figure 2. Co-occurences window

The next phase is the Textual Analysis, here we start with the identification of the segments, these are saved in two file: the former is Lista dei Segmenti (con indice IS) which contains the segments with their relative number of occurrences, number of elements forming the segment and the IS index; the latter, named Lista dei segmenti Significativi is a list of the significant segments. Then we analyze the specificity of our Corpus together with the computation of the TFIDF index, these data are saved in the same location under the name of: Vocabolario and Lessico. The last operation of our study is the Textual Analysis with the concordances and co-occurrences computation (Figure 2), the latter can be found in the file Cooccorrenze e collocazioni significative.

The results of these operations show that the system without knowing anything of the data submitted can retrieve meaningful information, such as the TFIDF index that shows how much a word is important in the document, or the co-occurrences which show the main concepts of the text thanks to the couple of words that recur all the time. Unfortunately TaLTac can not perform all the analysis on our data set due to the fact that it is in English and so our work is not totally complete. Even though our results are few, they are very significative.

3.2 GATE Analysis

Our first operation with this software is the initialization of ANNIE with Defaults, once all the Processing Resources are loaded we run ANNIE and start in sequence:

- Document Reset PR The document reset resource enables the document to be reset to its original state, by removing all the annotation sets and their contents;
- ANNIE English Tokeniser The tokeniser splits the text into very simple tokens such as numbers, punctuation and words of different types;
- ANNIE Gazetteer The role of the gazetteer is to identify entity names in the text based on lists;
- ANNIE Sentence Splitter The sentence splitter is a cascade of finite-state transducers which segments the text into sentences;

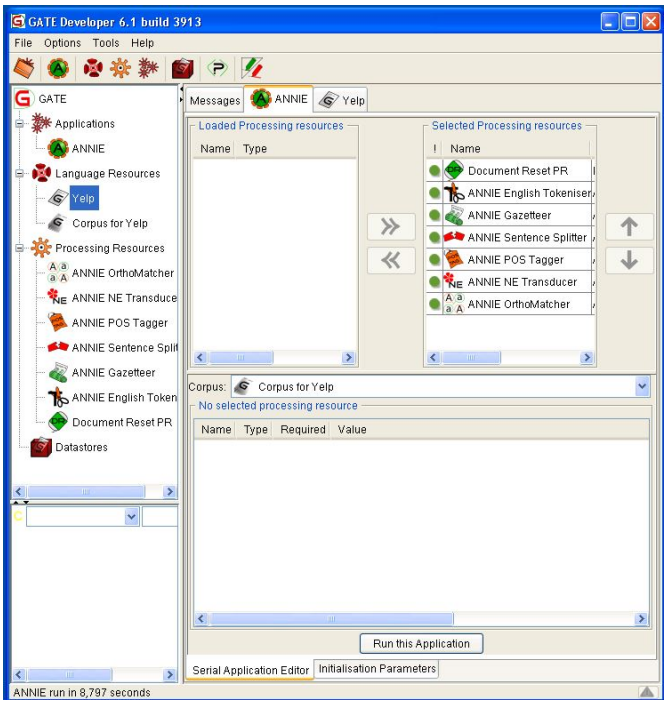


Figure 3. Annie Pipeline

- ANNIE POS Tagger The tagger produces a part-of-speech tag as an annotation on each word or symbol;
- ANNIE NE Transducer
- ANNIE OrthoMatcher The Orthomatcher module adds identity relations between named entities found by the semantic tagger, in order to perform coreference.

Thanks to these operations we can overcome TaLTac limits and run a Grammatical Tag on our data set. It must be said that our original data have been modified, indeed we are going to analyze only the review section due to the fact that in GATE the other information are meaningless.

4 Experimental Results

From our initial data set, made up of more than 5000 tuples, we extracted, as said, our 50 samples from which we obtained these results:

4.1 TaLTac results

Thanks to TaLTac functions we can say that our peculiar lexicon, shown in Figure 4, is composed by words with highest occurrences and TFIDF like: games, beer, good, great, food, place so it is reasonable to assume that the reviews analyzed talk about a place to eat food, drink beer or

Forma grafica	Occorrenze totali	Lunghezze	TFIDF	Forma grafica	Occorrenze totali	Lunghezze	TFIDF
was	83	03	3,37217	they	38	04	1,82534
s	62	01	2,72497	here	21	04	1,82342
for	71	03	2,61815	area	15	04	1,79898
games	34	05	2,46866	bar	24	03	1,79825
is	97	02	2,36757	get	20	03	1,78191
beer	24	04	2,19451	great	20	05	1,77350
are	34	03	2,10028	as	17	02	1,77066
to	132	02	2,07886	with	37	04	1,76399
t	37	01	2,06400	had	27	03	1,75748
good	37	04	2,06279	it	57	02	1,75409
selection	15	09	2,06025	place	44	05	1,75382
you	46	03	2,04468	out	27	03	1,72490
were	32	04	1,98688	I	131	01	1,71641
there	29	05	1,97350	fun	13	03	1,70398
that	57	04	1,95741	4.0	15	03	1,65914
-	25	01	1,93592	It	18	02	1,65544
Great	7	05	1,92191	The	52	03	1,65391
in	56	02	1,91502	from	24	04	1,64308
on	47	02	1,90879	wings	9	05	1,64125
food	39	04	1,87488	be	18	02	1,62180
1	26	01	1,86714	&	20	01	1,59426
of	91	02	1,84201	have	46	04	1,59094
1qCuOcksSHRv67OHovA	26	22	1,58947	but	48	03	1,57259
all	18	03	1,58031	can	12	03	1,55776
and	194	03	1,57490	5.0	10	03	1,55417

Figure 4. Peculiar lexicon by occurences

play some games and that the main audience thinks that it is a good place. Other meaningful data are the co-occurrences which are reported in Figure 2, these elements strengthen our thoughts about the place reviewed with expressions as: the bar, a good, a place, beer selection, the games. Last we took a meaningful word, food, and studied its concordances through all the Corpus (Figure 5), in order to understand the context of the lemma.

4.2 GATE results

Thanks to the Tokeniser we can distinguish spaces from words, and thanks to the Gazetteer and POS Tagger every word in our data set has a description. At the end of our analysis with this software we can assert the grammatical features of our data and we can give even more meaning to the analysis described in section 4.1.

5 Conclusions

Social networks analysis is an emerging trend among scholars and researchers in the last years. In literature, there are various instruments and project to achieve different kind of analysis, yet in our work, we focused on the analysis of the content of the text obtained from the social media. Through our investigation it was possible extracting the hot topics for the different information sources, as, in our case, the reviews analyzed. This study can be the starting point of further analysis on the domain of cybersecurity, through the detection of text containing dangerous messages, viral market advertising, thanks to an analysis of the feedback from the users or costumers, or information crawling.

ID Fr...	Intorno sinistro	Forma grafica	Intorno destro
fragm...	, and authentic . If you' re looking for good Irish	food	and a cold pint , you can' t go wrong at the Pour House
fragm...	worth seeking out . They have some of the best Irish	food	I' ve had in Pittsburgh- the colcannon is awesome and
fragm...	out of this world . If you' re not looking for Irish	food	, then try the grilled cheese- and make sure you ask
fragm...	there on a Saturday night with a mind to try the Irish	food	. Apparently , we were out of luck . I' ve always thought
fragm...	secrets of restaurant success is to actually stock	food	for people to eat . He told us before we ordered that
fragm...	were out " . At that point , realizing that the only	food	to be had in the place was what was crusted on the
fragm...	heoc96QXrTbecWVw933qhQ ǀ 2011-08-20 Best Irish	food	in the Burgh . Great bar food too . The service is
fragm...	2011-08-20 Best Irish food in the Burgh . Great bar	food	too . The service is maybe a bit surly and it' s not
fragm...	Excellent wings and sandwiches , generally good	food	otherwise , and fair prices-- nice casual place . The
fragm...	1qCuOcks5HRv67OHovAVpg tPUGLIDZLF7HrOC46NqT...	food	here can actually be a little hit and Miss , but I
fragm...	wonderful character and ambiance of this place , but the	food	was average at best . We were there on a Pens play-off
fragm...	2 people working all the tables in both rooms . Our	food	came to us cold and unimpressive at that . Our orders
fragm...	were pretty good , also . If you are looking for good	food	, Homestead better choices at Blue Dust or Tin Front
fragm...	beer selection , but it won' t be for a while . The	food	has always been average at best , and the pizza sub
fragm...	op2Gve4sAMQ4qEzq2Tad0g ǀ 2013-09-15 I' m reading o...	food	is really hit or Miss . I' ve only gone to Duke' s
fragm...	Miss . I' ve only gone to Duke' s one time , but the	food	was good . It wasn' t too busy so our service was very
fragm...	very attentive and it didn' t take long to get our	food	at all . The place is separated between a bar area

Figure 5. Concordances of the word "food"

References

[1] M. Albanese, A. D'acierno, V. Moscato, F. Persia, and A. Picariello. Modeling recommendation as a social choice problem. pages 329–332, 2010.

[2] F. Amato, M. Barbareschi, V. Casola, and A. Mazzeo. An fpga-based smart classifier for decision support systems. *Studies in Computational Intelligence*, 511:289–299, 2014.

[3] F. Amato, M. Barbareschi, V. Casola, A. Mazzeo, and S. Romano. Towards automatic generation of hardware classifiers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8286 LNCS(PART 2):125–132, 2013.

[4] Flora Amato, Antonino Mazzeo, Vincenzo Moscato, and Antonio Picariello. A framework for semantic interoperability over the cloud. In *Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on*, pages 1259–1264. IEEE, 2013.

[5] R. Aversa, B. Di Martino, and F. Moscato. Critical systems verification in metamorp(h)osy. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8696 LNCS:119–129, 2014.

[6] Walter Balzano and Maria Rosaria Del Sorbo. Cotracks: A new lossy compression schema for tracking logs data based on multiparametric segmentation. In *Data Compression, Communications and Processing (CCP), 2011 First International Conference on*, pages 168–171. IEEE, 2011.

[7] Walter Balzano and Maria Rosaria Del Sorbo. Setra: A smart framework for gps trajectories' segmentation. In *Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on*, pages 362–368. IEEE, 2014.

[8] Walter Balzano and Fabio Vitale. Dig-park: a smart parking availability searching method using v2v/v2i and dgp-class problem. In *31st International Conference on Advanced Information Networking and Applications Workshops 2017 Computer Science*, 2017.

[9] Mario Barbareschi, Alessandra De Benedictis, Antonino Mazzeo, and Antonino Vespoli. Providing mobile traffic analysis as-a-service: Design of a service-based infrastructure to offer high-accuracy traffic classifiers based on hardware accelerators. *Journal of Digital Information Management*, 13(4):257, 2015.

[10] Mario Barbareschi, Salvatore Del Prete, Francesco Gargiulo, Antonino Mazzeo, and Carlo Sansone. Decision tree-based multiple classifier systems: An fpga perspective. In *International Workshop on Multiple Classifier Systems*, pages 194–205. Springer, 2015.

[11] Mario Barbareschi, Antonino Mazzeo, and Antonino Vespoli. Malicious traffic analysis on mobile devices: a hardware solution. *International Journal of Big Data Intelligence*, 2(2):117–126, 2015.

[12] Alessandro Cilardo, Mario Barbareschi, and Antonino Mazzeo. Secure distribution infrastructure for hardware digital contents. *IET Computers & Digital Techniques*, 8(6):300–310, 2014.

[13] Enrico Coiera. *Guide to health informatics*. CRC press, 2015.

[14] Francesco Colace, Massimo De Santo, and Luca Greco. A probabilistic approach to tweets' sentiment classification. In *Affective computing and intelligent interaction (ACII), 2013 humaine association conference on*, pages 37–42. IEEE, 2013.

[15] Francesco Colace, Massimo De Santo, Luca Greco, Vincenzo Moscato, and Antonio Picariello. A collaborative user-centered framework for recommending items in online social networks. *Computers in Human Behavior*, 51:694–704, 2015.

[16] A. D'Acierno, V. Moscato, F. Persia, A. Picariello, and A. Penta. iwin: A summarizer system based on a semantic analysis of web documents. pages 162–169, 2012.

[17] Son Doan, Lisa Bastarache, Sergio Klimkowski, Joshua C Denny, and Hua Xu. Integrating existing natural language processing tools for medication extraction from discharge summaries. *Journal of the American Medical Informatics Association*, 17(5):528–531, 2010.

[18] Georg Fette, Maximilian Ertl, Anja Wörner, Peter Kluegl, Stefan Störk, and Frank Puppe. Information extraction from unstructured electronic health records and integration into a data warehouse. In *GI-Jahrestagung*, pages 1237–1251, 2012.

[19] Amit X Garg, Neill KJ Adhikari, Heather McDonald, M Patricia Rosas-Arellano, PJ Devereaux, Joseph Beyene, Justina Sam, and R Brian Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, 293(10):1223–1238, 2005.

[20] U Kang, Duen Horng Chau, and Christos Faloutsos. Pegasus: Mining billion-scale graphs in the cloud. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5341–5344. IEEE, 2012.

[21] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

[22] Randolph A Miller. Medical diagnostic decision support systemspast, present, and future. *Journal of the American Medical Informatics Association*, 1(1):8–27, 1994.

[23] F. Moscato. Model driven engineering and verification of composite cloud services in metamorp(h)osy. pages 635–640, 2014.

[24] F. Moscato. Exploiting model profiles in requirements verification of cloud systems. *International Journal of High Performance Computing and Networking*, 8(3):259–274, 2015.

[25] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics*, pages 643–674. Springer, 2014.

[26] The Apache Hadoop project. Apache hadoop.

[27] The GATE project team. Gate.

[28] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[29] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't, 1996.

[30] Adolfo Morrone Sergio Bolasco, Francesco Baiocchi. Taltac.

[31] Samson W Tu, James R Campbell, Julie Glasgow, Mark A Nyman, Robert McClure, James McClay, Craig Parker, Karen M Hrabak, David Berg, Tony Weida, et al. The sage guideline model: achievements and overview. *Journal of the American Medical Informatics Association*, 14(5):589–598, 2007.

[32] Manuela Veloso, Jaime Carbonell, Alicia Perez, Daniel Borrajo, Eugene Fink, and Jim Blythe. Integrating planning and learning: The prodigy architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):81–120, 1995.

[33] The Free Encyclopedia Wikipedia. Yelp.