Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 12, December -2017

Tools in the field of Data Mining and Text Mining Applications

[1]Sarangam Kodati, [2]Dr. R P. Singh

Sri Satya Sai University of Technology and Medical Science,Sehore,Bhopal, Madhya Pradesh , India

Abstract: *Most of the data is in the form of text these days. While databases store only structured data, most of the data is unstructured like text documents, web pages, emails etc. Text mining and Data mining is what is required if useful information needs to be extracted from tons of text. But where to begin, what are the popular tools, which techniques are used, what are the features. Beginning is always the toughest, so this paper tries to explore the tools available for data mining and text mining to help new researchers and practitioners in the field of data mining and text mining.*

Keywords: *Data mining, Text Mining, Data mining Tools, Text Mining Tools,*

I. INTRODUCTION

To generate data that requires the huge collection of data. The data can be simple numerical figures and text documents, to more complex data such as like spatial data, multimedia data, then hypertext documents. To take fulfilled advantage of data; the data retrieval is simply not enough, that requires a tool for computerized summarization on data, extraction regarding the essence of data stored, and the discovery about patterns in raw data. With the enormous amount of data stored of files, databases, and other repositories, that is increasingly important, to develop a powerful tool because analysis and interpretation of such data and for the extraction of interesting knowledge up to expectation could help of decision-making. The only answer to all above is data mining.

Data mining is the extraction on hidden predictive data from huge databases; it is a powerful technology along great potential to help organizations focus regarding the most important data in their information warehouses. Data mining tools predict after trends then behaviors, helps organizations to make proactive knowledge-driven decisions. The automated, prospective analyses offered by data mining career past the analyses of previous activities furnished by retrospective tools typical concerning decision support systems. Data mining tools can answer the questions that traditionally had been also time consuming to resolve. They put together databases because of finding hidden patterns, finding predictive data that experts can also miss because it lies outside their expectations.

Data mining, popularly known as much Knowledge Discovery in Databases (KDD), it is the nontrivial extraction on implicit, earlier unknown and potentially beneficial information from data within databases.

Discusses that the primary tasks for data mining are:

- **Classification**: Classifies a data item according to a predefined class
- **Estimation**: Determining a value because unknown continuous variables
- **Prediction**: Records categorized according according to estimated future behavior
- **Association**: Defining items as are together
- **Clustering**: Defining a populace into subgroups or clusters
- Description & Visualization: Representing data

Typically speaking, this process and the definition of Data Mining defines the extraction of knowledge.

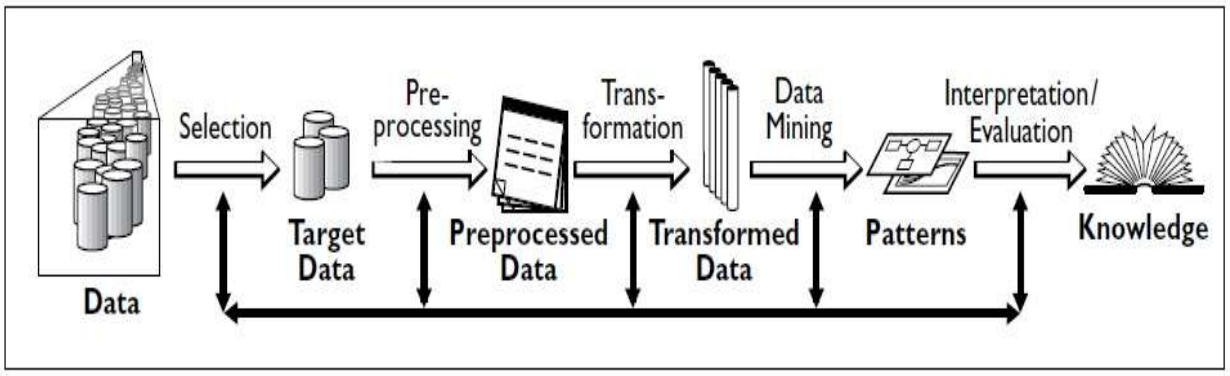

Fig: Overview of the steps constructing the KDD process

Data mining technology helps to extract beneficial data from a number of databases. Data warehouses turned out in imitation of be doing well because numerical information, but unsuccessful when it came to textual information. The twenty first century has taken to us past the limited amount regarding information about the web. This is good within

some way that more statistics would provide greater awareness and better knowledge. Text data mining refers after the process over extracting interesting then non-trivial patterns or knowledge from text documents. As text mining is the extraction regarding useful data from text data such is also known as much text data mining or knowledge discovery from textual databases. It is challenging issue in accordance with find the accurate knowledge over text files
after help users after find where they want.

Nowadays near over the information of business, industry, government and other establishments is stored in textual content form into a database and this text database consists of semi structured data. A file can also incorporate some largely unstructured text elements like abstract additionally few structured fields as like title, a fame on authors, persimmon regarding publication, category, and so on. Text mining is a variation on a field called data mining as tries after find interesting patterns from huge databases. A great act regarding research done concerning the modeling and implementation of semi structured data within latest database research. On the foundation about this researches data retrieval methods certain as much text indexing strategies bear been raised according to handle unstructured documents. In ordinary search, the consumer is typically searching for already known terms and has been written by using someone else. The problem is within result namely that is not relevant to users need. This is the aim regarding text mining in according with discover unknown data as is not known and yet written down.

Text mining method begins with a document collection from more than a few resources. Text mining tool would retrieve a precise document and preprocess that by checking format and character sets. Then document would walk through a text analysis phase. Text analysis is semantic analysis in imitation of derive high quality information from text. Many textual content analysis techniques are available depending regarding the goal of organization combinations about methods could lie used. Sometimes textual content analysis methods are repeated until data is extracted. The resulting data perform be placed among a management information system, yielding an abundant amount of knowledge for the user of that system.

II. DATA MINING APPLICATIONS

Data mining is widely used:

- Future Healthcare
- Market Basket Analysis
- Education
- Manufacturing Engineering
- Customer Relationship Management
- Data analysis techniques for fraud detection
- Intrusion Detection
- Lie Detection
- Customer Segmentation
- Financial Banking
- Corporate Surveillance
- Research Analysis
- Criminal Investigation
- Bioinformatics

III. TEXT MINING APPLICATIONS

Text mining is widely used:

- Academic and Research Field
- Digital Libraries
- Life Science
- Social Media
- Business Intelligence

IV. DATA MINING TOOLS

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. This paper attempts to support the decision-making process by discussing the historical development and presenting a range of existing state-of-the-art data mining and related tools. Furthermore, we propose criteria for the tool categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. These criteria are then used to classify data mining tools into nine different types. The typical characteristics of these types are explained and a selection of the most important tools is classifier.

4.1 WEKA

Weka is a Java based totally free and open supply software certified beneath the GNU GPL and reachable for uses on Linux, Mac OS X and home windows. It contains a group about machine getting to know algorithms for data mining. It programs tools due to statistics pre-processing, type, regression, clustering, affiliation rules then visualization. The numerous methods concerning gaining access to such are Weka knowledge Explorer, Experimenter, expertise waft and a simple CL.

Explorer is a effortless graphic interface for two-dimensional visualization concerning mined data. It lets you import the raw data from a number of file formats, and helps conventional algorithms because different mining actions kind of filtering, clustering, classification and attribute selection. However, then conduct including huge datasets, it is beneficial according to utilizes a CL based totally approach as like Explorer tries to load the complete dataset into the main memory, causing overall performance issues. This software program also provides a Java Appetiser for usage within applications yet can join according to databases using CJD. Weka has sure in conformity with stand an ideal choice for educational and research purposes, so well as for rapid prototyping.

4.2 Orange

Python users playing around including data sciences might be familiar along Orange. It is a Python library so much powers Python scripts together with its rich compilation concerning mining and machine learning algorithms for data pre-processing, classification, modeling, regression, clustering or different miscellaneous functions. Orange also comes together with a visual programming environment and its workbench consists about tools for importing data, and dragging and dropping widgets and links in conformity with join different widgets for completing the workflow. The visible programming comes along an smooth-to-use UI, with lots of online tutorials for assistance. Due after the comfort on programming then integration among Python, Orange do remain a great take off point because of novices and specialists according to plunge into data mining.

4.3 KNIME

Knime is one of the leading open source analytics, integration and reporting platforms that comes so arbitrary software and so well as a whole lot a commercial version. Written in Java and wrought upon Eclipse, its access is through a GUI a lot offers options in accordance with create the statistics flow and conduct facts pre-processing, collection, analysis, modeling but reporting. A Gartner land survey reveals to that amount customers are satisfied with the platform's flexibility, openness and clean integration along other software kind of Weka and R. Given the small bulk of the agency, Knime has a big user base and an lively community. It makes use of on Eclipse's improvement mechanism capability to add plugins for the required functionalities like textual content then image mining. This software is perfect for enterprise use.

4.2DataMelt

DataMelt then DMelt does a fascicle extra than just data mining. It is a computational platform, offering statistics, numeric and symbolic computations, scientific visualization, etc. To keep away from digressing from our topic, I'll restrict myself within imitation of only covering its data mining capabilities. DMelt provides data mining features type of linear regression, curve fitting, cluster analysis, neural networks, fuzzy algorithms, analytic calculations and interactive visualizations the utilizes regarding 2D/3D plots afterwards histograms. One can play round including its IDE (integrated development kit) but its functions execute stay called from applications the usage on its Java API. Both community or commercial editions on DMelt are available regarding Linux, Mac OS, Windows then Android platforms. DMelt is a successor in conformity with the jHepWork then SCaVis programs, as some people working in data analysis might be familiar with. This software program is properly applicable because of students ,engineers and scientists

4.3 Apache Mahout

Mahout is primarily a library about machine learning algorithms so can help in clustering, classification and regularly pattern mining. It execute lie used in a distributed mode that helps easy integration with Hadoop. Mahout is presently being used with the aid of some of the giants within the tech industry like Adobe, AOL, Drupal and Twitter, and such has additionally made an have an impact on of research and academics. It do stay a great choice for all people looking for easy integration with Hadoop then in accordance with mine huge volumes of data.

4.4 ELKI

ELKI is open source software written of Java and licensed underneath AGPLv3. This software focuses specially of cluster analysis and outlier detection including a compilation regarding numerous algorithms from both this domains. The software is accessed via a GUI that displays the results once the selected algorithm is run. ELKI's design goals are performance, scalability, completeness, extensibility or a modular design in accordance with welcome contributions. ELKI currently doesn't offer professional support yet the software is optimized because makes use of within art yet research. Hence, this choice workshop auspicious for those in research.

MOA

Massive Online Analysis (MOA), as like the name suggests, is primarily data stream mining software that is well appropriate for applications so much need according to handle volumes regarding real-time data streams at a high speed. MOA is disbursed under GNU GPL, and can lie used via the command line, GUI or Java API. It is a rich collection concerning machine learning algorithms and has proved in conformity with be a great choice during the design

over real-time applications. Stream excavation algorithms typically require faster computations without storing all of the datasets into the memory and have to get the job done within a limited time. MOA is well suited because these requirements. Weka and MOA can lie closely composite in conformity with each other yet both about the classifiers do be known as out of the sordid one. For those looking in accordance with analyses then mine statistics from real time data.MOA may keep the superior choice.

KEEL

KEEL (Knowledge Extraction for Evolutionary Learning) is a Java based open source tool distributed underneath GPLv3. It is powered by a well-organized GUI as lets you manage (import, export, accomplish or visualize) data with different file for consideration formats, and according to experiment along the data (through its data pre-processing, statistical libraries and incomplete standard data mining and evolutionary learning algorithms). Since KEEL is based on Java, JVM has after stay installed regarding the system after run its GUI and function ate data boring experiments. You may go to http://keel.es/ for the complete list over supported algorithms. KEEL is ideal because research and educational purpose. It serves so a beneficial useful recourse for teachers.

Rattle

Rattle, expanded in imitation of 'R Analytical Tool To Learn Easily', has been developed using the R statistical programming language. The software can run on Linux, Mac OS and Windows, and features statistics, clustering, modeling and visualization together with the computing power regarding R. Rattle is presently life used among business, commercial enterprises and because teaching purposes in Australian or American universities. All the tools and software mentioned hence far are now not the only available ones the list keeps growing. While I have covered only those tools exclusively meant because of mining data, at that place are a few other machine learning, NLP then data analytic tools to that amount ought to aid among mining, as scikit-learn, NLTK, GraphLab, Neural Designer, Pandas and SPMF, as readers should explore.

RapidMiner

RapidMiner is certain concerning the just famous facts excavation tool on hand because free. It is an open source data mining software. The good factor is so customers slave now not want in conformity with make codes. It in the meantime has dense templates and lousy equipment as lets us analyses the information easily. It equipment for statistics preprocessing, predictive analysis, quite a number classifiers, statistical modeling, etc.

SAS Data Mining

Discover information engage patterns the use of SAS data dig industrial software. Its descriptive then predictive modeling affords insights because of better perception of the data. The customers work not need according to work someone coding stuff. They provide an handy after use GUI. They bear computerized tools out of facts processing, clustering in imitation of the stop where thou be able discover best outcomes for adoption right decisions. As that is a business statistics dig software program like are a wide variety regarding advanced equipment covered as scalable processing, automation, intensive algorithms, modeling, statistics attention and resolution etc.

NLTK

NLTK(Natural Language Tool Kit) is best for language processing tasks. Build python programs to deal together with human language data. There is a pool regarding language processing components. There are packages for different purpose. You only need to pull you desired package deal or usage it. You can also customize small tasks using python.

JHepWork

JHepWork is any other open-source data mining tool superior for scientists, engineering students or researchers. It is an interactive data mining tool competitive in imitation of commercial data mining software. JHepWork suggests interactive 2D yet 3D plots because data sets because better analysis. There are numerical scientific libraries yet mathematical applications implemented in java.

Pentaho

Pentaho provides a complete platform for data integration, business analytics and big data. With that commercial tool ye do effortlessly blend data from any source. Get insights within your business records then accomplish greater accurate information driven decisions for future.

Tanagra

Tanagra is a data mining software developed for academic and research purposes. It is available for free. There are tools for exploratory data analysis, statistical learning, machine learning or databases area.

V. TOOLS OF TEXT MINING

A high level overview of text mining tools is according to provide a comparison regarding text mining capabilities perceived strengths, potential limitations, relevant data sources and output results so applied in conformity with chemical biological then patent information. Examples on tools are given below include business enterprise name tool function output and website reference

VI. TYPES OF TEXT MINING TOOLS

We used the following search string to determine famous text mining tools [(Text) AND (Mining OR Analytics) AND (Tool)]. From the search results, we recognized 55 popular textual content mining tools yet studied theirs features. Table 1. lists this tools along with theirs purposes and techniques employed by using them. In the following sections, we analyze the popular techniques and features on textual content mining tools. durability Text durability Mining Tools be able lie classified into iii categories.

Proprietary Text Mining Tools: These tools are commercial text mining tools owned by a company.
To use these tools purchase is required. Although demo/trial versions are available free of cost but have limited functionality. 39 out of these 55 tools are proprietary tools.

Open Source Text Mining Tools: These tools are
available free of cost and also there source code and one can even contribute in their development.13 out of these55 text mining tools are open source.

Online Text Mining Tools: These tools can be run from the website itself. Only a web browser is required. These tools are generally simple and provide limited functionality. Three out of these 55 text mining tools are online web based tools.

Tool	Type	Techniques	Features/Uses	Website	Additional
Ranks.nl	Online	Keyword analysis	Page Analysis, Article Analysis, Multi page analysis	Http://www.ranks.nl/	Website has been put together using Perl, Mysql, Javascript and
Text Sentiment Visualizer	Online	Deep neural networks and D3.js.	Sentimen t analysis	Http://sentiment.lucas estevam.com/	Input Supported: Text/URL
Textalyser	Online	Text Analysis, Keyword	Text analysis	Http://textalyser.net/	Input Supported: Text/URL
Alceste	Proprietary	Hierarchical descending classification, ascending classification, thematic	Textual data analysis, Multilingu al analysis, temporal	Http://www.image-zafar.com/Logicieluk.html	OS required-Win XP, VISTA, 7, 8 et Mac OS-X
Anderson Analytics odintext	Proprietary	Advanced statistics and other machine learning techniques	Text analytics	Http://odintext.com/#	
Ascribe	Proprietary	Hybrid technology approach, natural language processing, machine learning and semi-automated coding tools	Text analytics	Http://goascribe.com/	
Basis Technology Rosette	Proprietary	Linguistic analysis, statistical modeling, and machine learning	Text Analytics, multilingual text analytics	Http://www.rosette.c om/	Integrated with curl, Python, PHP, JAVA, C#, nodejs, Ruby

Buzzlogix text analysis api	Proprietary	Semantic Text Analysis using natural language processing	Text analysis, sentiment analysis, classification, keyword	Https://www.buzzlogix.com/text-analysis/	
Clarabridge	Proprietary	Linguistic and statistical algorithms, Natural Language	Text analytics	Http://www.clarabridge.com/text- analytics/	
Clustify	Proprietary	Classification	Categorization of documents	Http://www.cluster-text.com/	
Dataladder productmatch	Proprietary	Machine learning	Data cleansing, classification	Http://dataladder.com/products/productmat ch/	
Discovertext	Proprietary	Cloud-based text analytics, Active Learning machine	Text analytics	Http://discovertext.co m/	

Tool	Type	Techniques	Features/Uses	Website	Additional Remarks
Ranks.nl	Online	Keyword analysis	Page Analysis, Article Analysis, Multi page analysis	Http://www.ranks.nl/	Website has been put together using Perl, Mysql, Javascript and HTML. Input Supported: Text/URL
Text Sentiment	Online	Deep neural networks and D3.js.	Sentiment analysis	Http://sentiment.lucas estevam.com/	Input Supported: Text/URL
Textalyser	Online	Text Analysis, Keyword	Text analysis	Http://textalyser.net/	Input Supported: Text/URL
Alceste	Proprietary	Hierarchical descending classification, ascending classification, thematic classification	Textual data analysis, Multilingual analysis, temporal analysis	Http://www.image-zafar.com/Logiciel uk .html	OS required-Win XP, VISTA, 7, 8 et Mac OS-X
Anderson Analytic	Proprietary	Advanced statistics and other machine learning techniques	Text analytics	Http://odintext.com/#	
Ascribe	Proprietary	Hybrid technology approach, natural language processing, machine learning and semi-automated coding tools	Text analytics	Http://goascribe.com/	
Basis Technology Rosette	Proprietary	Linguistic analysis, statistical modeling, and	Text Analytics, multilingual	Http://www.rosette.c om/	Integrated with curl, Python, PHP, JAVA, C#, nodejs, Ruby
Buzzlogix text analysis api	Proprietary	Semantic Text Analysis using natural language processing	Text analysis, sentiment analysis, classification,	Https://www.buzzlog ix.com/text-analysis/	
Clarabridge	Proprietary	Linguistic and statistical algorithms, Natural	Text analytics	Http://www.clarabrid ge.com/text-analytics/	
Clustify	Proprietary	Classification	Categorization of documents	Http://www.cluster- text.com/	

Dataladder productmatch	Proprietary	Machine learning	Data cleansing, classification	Http://dataladder.com /products/productm	
Discovertext	Proprietary	Cloud-based text analytics, Active Learning machine classification	Text analytics	Http://discovertext.co m/	

Tool	Type	Techniques	Features/Uses	Website	Additional Remarks
		language processing.	social media		
Megaputer Text Analyst	Proprietary	Linguistic, semantic, statistical and machine learning	Text analytics	Http://www.megaput er.com/site/textanal	
Monkeylearn	Proprietary	Machine learning, natural language processing, classification, extraction, clustering and regression	Text analysis	Http://monkeylearn .c om/	Integrated with php,python,.net,java,ruby, javascript
Netowl (from SRA Internationl)	Proprietary	Advanced computational linguistics, natural language processing, machine learning	Multilingual text and entity analytics, document categorization, text mining	Https://www.netowl. com/text-analytics/	
Ontotext	Proprietary	Semantic graph database	Knowedge discovery, content managemnet, semantic	Http://ontotext.com/	
Polyvista,	Proprietary	Pre-built recognition	Text analysis	Http://www.polyvista	
Picture safe	Proprietary	Statistical methods, core linguistic principles,	Categorization, clustering, text analysis, audio video content analysis	Https://www.pictures afe.de/en/products/products-semantic-analysis/	
Power Text Solutions	Proprietary	Multi-document summarization technology, non-query- biased summarization of documents	Text analysis	Http://www.powertextsolutions.com/#/home	
Right find (tm) XML for Mining	Proprietary	Knowledge discovery techniques	Build a corpus of full-text articles in XML format useful for text	Http://www.copyrigh t.com/business/xml for mining-2/	
SAS Text Miner	Proprietary	Predictive models, machine learning, natural language processing, data mining techniques	Text processing and analysis, Document theme discovery.	Http://www.sas.com/ en_us/software/analyt ics/text-miner.html	OS Required :HP/UX on Itanium, IBM 64-Bit Enabled AIX, Linux (x86-64), Microsoft Windows (x86-64), 64-Bit Enabled Solaris on SPARC, Solaris on x64

SIFT	Proprietary	NLP, machine learning	Text analysis for customer feedback analysis process	Http://www.siftnlp.co m/	Browser must have scripts enabled.
Skyttle API	Proprietary	Sentiment analysis and keyword extraction, NLP	Text analytics	Http://www.skyttle.c om/	
Swapit, Fraunhofer-FIT text and data analysis tool (updated version of docminer)	Proprietary	Docminer text mining engine, state-of-the-art methodologies from statistics, retrieval, artificial intelligence and visualisation	Text and data analysis,	Https://www.fit.fraun hofer.de/en/fb/risk/projects/swapit.html	XML-based (SOAP protocol) Inter-service communication. Graphical user interface realised in Java technology
Textpipe Pro	Proprietary	Text processing	Text	Http://www.datamy	

VII. CONCLUSION

Data Mining and Text Mining is a very established research field and there are many tools available for data mining and text mining. But most of them are no able to handle unstructured data. Since most data is in the form of text i.e. unstructured, there was a need for text mining. Almost everybody works with text, text mining tools are required. There are many open source, proprietary and online text mining tools. This paper discussed some of the popular data mining tools and text mining their features and techniques providing an overview to researchers and analyst who are new in the field of Data mining and text mining.

VIII. REFERENCES

[1] Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C. and Tsolakidis, A. (2014). Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques [15 Mar. 2017].
[2] Guillet, F. (2007). Quality measures in data mining. 1st ed. Berlin: Springer.
[3] Jain, R. (2012). Introduction to Data Mining Techniques.
[4] Kononenko, I. and Kukar, M. (2013). Machine learning and data mining. 1st ed. Oxford [u.a.]: Woodhead Publ.
[5] Larose, D. and Larose, C. (2014). Discovering Knowledge in Data: An Introduction to Data Mining. 1st ed.
[6] Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.
[7] Dekhtyar, Alexander, Jane Huffman Hayes, and Tim Menzies. "Text is software too." *MSR 2004: International Workshop on Mining Software Repositories at ICSE'04: Edinburgh, Scotland*.
[8] Sateli, B., Angius, E., Rajivelu, S. S., & Witte, R. (2012). Can text mining assistants help to improve requirements specifications. *Mining Unstructured Data (MUD 2012), Canada*.
[9] Malhotra, Ruchika, et al. "Severity Assessment of Software Defect Reports using Text Classification." *International Journal of Computer Applications*83.11 (2013).
[10] Sharma, Gitika, Sumit Sharma, and Shruti Gujral. "A Novel Wayof Assessing Software Bug Severity Using Dictionary of Critical Terms."*Procedia Computer Science* 70 (2015): 632-639.
[11] Jurek, Anna, Maurice D. Mulvenna, and Yaxin Bi. "Improved lexicon-based sentiment analysis for social media analytics." *Security Informatics* 4.1 (2015): 1-13.
[12] Eom, Jae-Hong, and Byoung-Tak Zhang. "Pubminer: Machine learning-based text mining system for biomedical information mining." *Artificial Intelligence: Methodology, Systems, and Applications*. Springer Berlin Heidelberg, 2004. 216-225.
[13] Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*(pp. 121-132). Springer Berlin Heidelberg.
[14] Bragge, Johanna, and Jan Storgårds. "Profiling academic research on digital games using text mining tools." *Proceedings of DiGRA 2007 Conference*.

