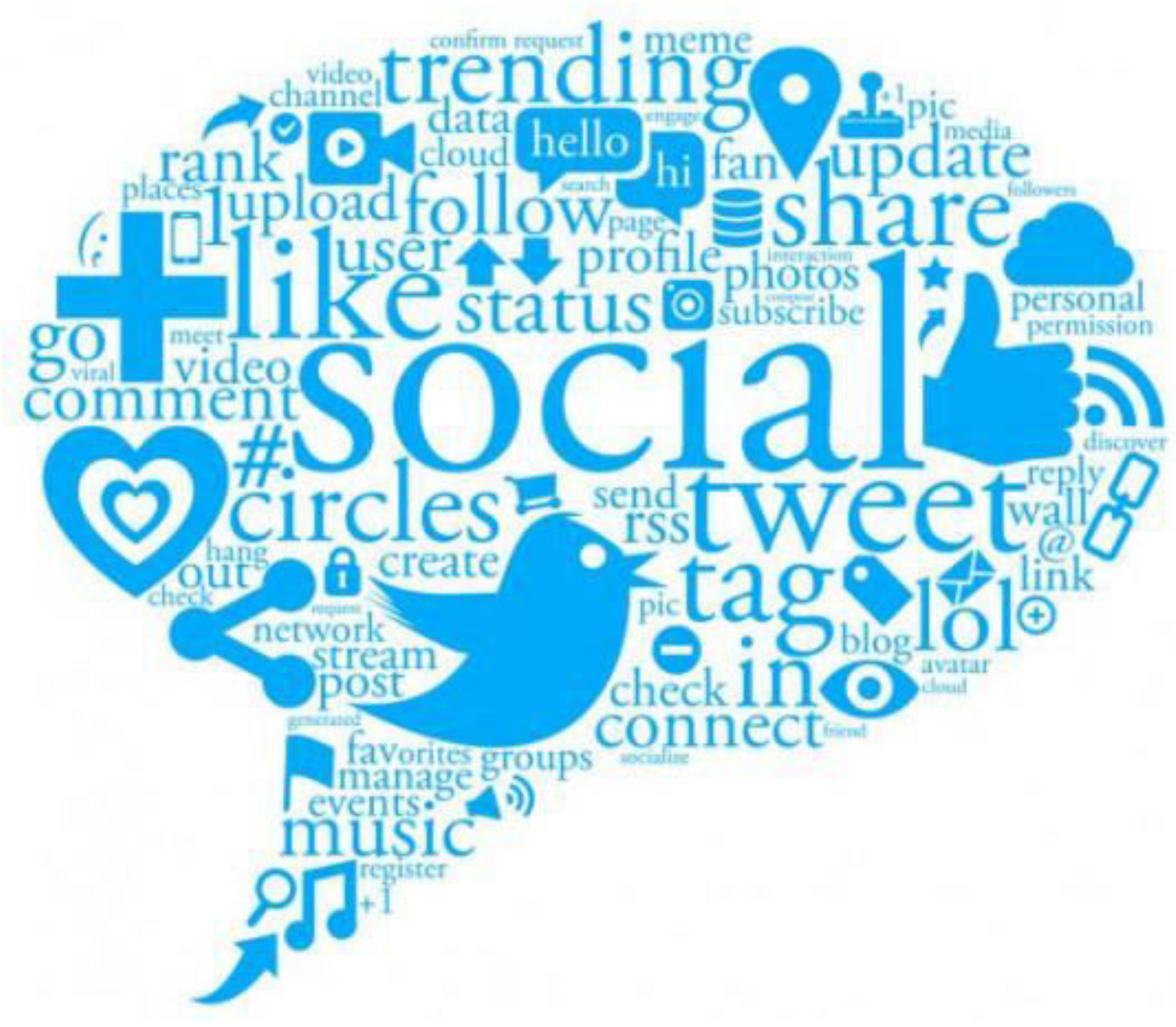
Research manual



Student: Ignas Usinskas

Student ID: C00166783

Supervisor: Greg Doyle

Date: 11/11/2015

Table of Contents

1. Introduction

In this document I will be talking about sentiment analysis – what it is, who worked in sentiment analysis area, what types of sentiment analysis are used today. For this particular project I will look into existing technologies, what people use to implement sentiment analysis and how I can implement it. I will compare many different ways of implementing it and will decide which way is the best and why. We will look into many databases and decide which one suits this project best. Research will be done and displayed for similar applications to this project that exist right now. By the end of this research document I will have come to the conclusion of how to implement sentiment analysis for this project and what technologies would be best suited for achieving a working and reliable application.

2. Sentiment analysis

Sentiment analysis is identification and extraction of subjective information from source materials using natural language processing, text analysing and computational linguistics. Sentiment analysis is extensively applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

Sentiment analysis includes a progression of different techniques to decide the state of mind of a speaker or a writer as for some topic or general relevant extremity of a document. This is extremely valuable for organizations as sentiment analysis gives and overall assessment of their item taking into account individuals' sentiments.

A primary use for sentiment analysis is to classify the polarity of given text at the document, sentence or feature/aspect level and determine whether the expressed opinion is positive, negative or neutral.

Turney[1] and Pang[2] in early work in sentiment analysis area applied different methods for recognising extremity of product and movie audits respectively. A classification of document's extremity of multi-way scale was endeavoured by Pang[3] and Snyder[4].

To determine whether a given text is positive, negative or neutral in sentiment analysis a scaling system is used. Particular words which have positive, negative or neutral meaning have a rating associated with them from -10 to +10. When text is analysed using natural language processing, the subsequent concepts are analysed for an understanding of these words and how they relate to the given text. Every concept is given a score based on the way sentiment words identify with the concept and their related score. After scores have been given, an overall score of positive, negative or neutral is associated to the given content. Figure 1 below shows steps involved in sentiment analysis.

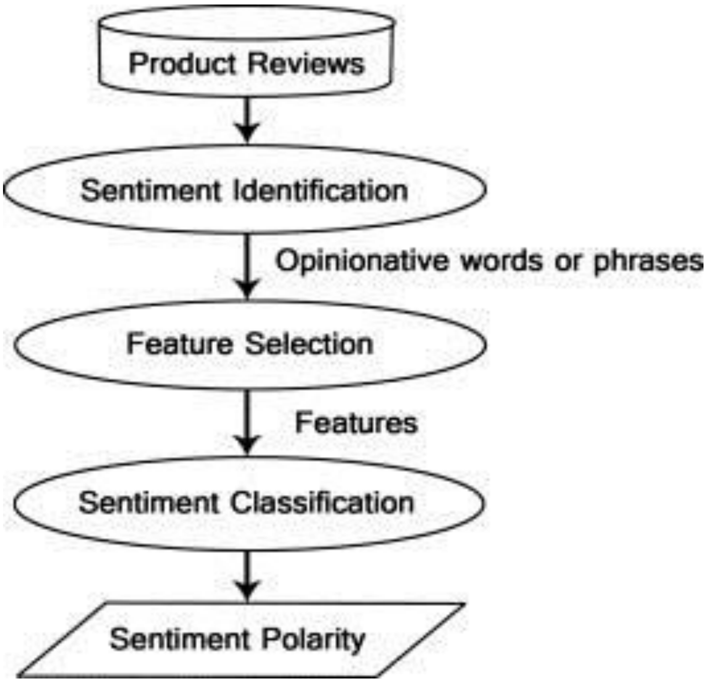


Figure1. Steps involved in sentiment analysis[5].

2.1 Types of sentiment analysis

- **Subjectivity/objectivity identification** – This type of sentiment analysis is defined as classifying a given text into one of two classes: objective or subjective. This type can sometimes be more difficult than polarity classification. Subjectivity of words and phrases can depend on their context and an objective document can contain subjective sentences[6].
- **Feature/aspect-based sentiment analysis** – Refers to determining the opinions or sentiments expressed on different features or aspects of entities such as cell phones, cars or buildings. A feature or an aspect is an attribute or a component of an entity, for example: the screen of a cell phone, cosmetic condition of a car or a colour of a building. The advantage of this type of sentiment analysis is the possibility to capture distinctions about objects of interest. Different features can produce different sentiment analysis results, for example: a car can be fast but require a lot of petrol. This problem associates several sub-problems such as identifying relevant entities, extracting their features/aspects and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral[7].

2.2 Existing approaches

There are four main existing approaches to sentiment analysis:

1. **Keyword spotting** – is the most naïve but also the most popular approach because of its accessibility and economy. Text is allocated into effect categories based on the presence of reasonably obvious affect words for example: happy, sad, afraid and bored. This approach has a shortcoming in two areas: poor realisation of affect when negation is

included and dependence on surface features. In the first place shortcoming identifies with distinctive phrasing of given content for instance: this approach can accurately group the sentence “today was a decent day” as the day being good, but it will probably fall flat on a sentence like “today wasn’t a decent day by any stretch of the imagination”. Second shortcoming depends on the vicinity of evident affect words that are only surface features of the content. In normal speech and text a lot of sentences bring affect through underlying meaning rather than affect adjectives. For instance, the sentence “My car was stolen and insurance agency will not repay me for it” clearly invokes forceful feelings, yet utilises no affect words thus cannot be classified utilising a keyword spotting approach[9].

2. **Lexical affinity** – is a bit more practical than keyword spotting, as this method does not just detect obvious affect words, it assigns arbitrary words a probabilistic ‘affinity’ for a particular emotion. For example, the word ‘accident’ can be assigned a 75% probability that it indicates a negative effect, as in ‘car accident’ or ‘hurt by accident’. These probabilities are mostly learnt using linguistic corpora. Even though this approach usually outperforms keyword spotting approach, there are two main problems. First, this approach operates only on the word-level and because of that it can easily be tricked by sentences such as “I avoided an accident” and “I met my girlfriend by accident”. Second, the probabilities of lexical affinity are usually based towards a particular genre text. This makes it difficult to develop a reusable, domain independent model[9].
3. **Statistical methods** – Bayesian inference method supports vector machine and artificial neuron network, this method is popular for affect classification of text. It is possible for a system to learn affective valence of affect keywords, punctuation and word co-occurrence frequencies by feeding a machine learning algorithm a large corpus of affectively annotated texts. Unfortunately in general, traditional statistical methods are weak, for cases except with clear affect keywords - other lexical or co-event components in a statistical model have minimal prescient esteem exclusively. As a result this approach provides acceptable accuracy only when given a sufficiently large text input. This prompts the content on the page or section level to be adequately classified, while this approach does not work as well on smaller content, for example, sentences or provisions[9].
4. **Concept-level approach** – this approach focuses on semantic analysis of text through the use of web ontologies or semantic networks, which allow the collection of conceptual and affective information related to natural language opinions. This approach steps away from blind usage of keywords and word co-occurrence counts by relying on large semantic knowledge bases and the implicit features associated with the natural language concepts. Concept-based approaches have advantage over purely syntactical techniques by being able to detect sentiments that are expressed in a subtle manner, for instance through the examination of concepts that do not expressly pass any feeling, yet are certainly connected to different concepts that do as such[9].

2.3 Evaluation

The accuracy of a sentiment analysis system is decided by how well it agrees with human judgments. This is commonly measured by precision and recall which is based on understanding and measure of relevance. That being said, according to research, human rates generally agree only 79% of the time. This means that if a program is 70% accurate – it is doing nearly as well as a human even if such accuracy does not sound impressive. Regardless of the fact that a program was 100% right of the time, people would still disagree with it around 20% of the time, because humans disagree that much about any answer[10].

2.4 Sentiment analysis demand

The rise of social media such as blogs, forums and online discussions in social networks such as Facebook and Twitter has increased interest in sentiment analysis. With the proliferation of reviews, ratings, recommendations and other forms of online expressions, online opinion has turned into a currency for businesses that are looking to market their products, recognize new open doors and deal with their notorieties.

The demand for sentiment analysis is rising and as such several research teams in universities around the globe are focusing on understanding the dynamics of sentiment in e-communities through sentiment analysis. For example the CyberEmotions project recently identified the role of negative emotions in driving social networks discussions[11].

Sentiment analysis is very complicated. The fact that humans often disagree on the sentiment of a given text exhibits how big of a task it is for computers to get this right. Cultural factors, linguistic nuances and different contexts make it difficult to a great degree to turn a string of written content to a basic pro or con sentiment. Since it is so troublesome – most sentiment analysis algorithms use straightforward terms to express sentiment about a product or a service and that is the place where the issue lies.

3. Technologies used

There are many ways sentiment analysis can be implemented in many different languages, first I will discuss how to implement it using python and machine learning algorithm and latter I will discuss other good alternatives and will determine which language and approach is best suited for this project.

3.1 Python

Python is a general-purpose, high-level programming language which is widely utilised[12][13]. It is anything but difficult to program and comprehend. Its outline reasoning stresses the simplicity of reading code and its linguistic structure permits software engineers to express ideas in less lines of code than it would be conceivable in different languages, for example, C++ or Java[14].

Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It also has a dynamic type system and automatic memory management and also has a large comprehensive standard library. Python code can be executed on wide variety of systems because its interpreters are available for installation on many operating systems[14].

3.2 Python history

Python was created in the late 1980s[15] and its implementation was started in December in 1989[16] by Guido van Rossum at CWI in the Netherlands. Python was developed as a successor language to the ABC language capable of execution handling and interface with the Amoeba operating system.

Python 2.0 was released on 16th October 2000 with many new major features such as a cycle-detecting garbage collector and support for Unicode.
Python 3.0 was released on 3rd December 2008 witch is a major, backwards incompatible release[17].

3.3 Python features

Python is a multi-paradigm programming language which supports object-oriented programming and structured programming fully. It has a number of language features which support functional programming and aspect-oriented programming. Many other paradigms are supported using extensions such as design by contract and logic programming[14].

For memory management, python uses dynamic typing, combination of referring counting and a cycle-detecting garbage collector for memory management. Python has a very important feature called dynamic name resolution which binds method and variable names during program execution[14].

Python has four most popular data types:

1. **List** – is a most versatile available in this programming language. This data type can be written as a list of comma-separated values between square brackets. Important thing to note about the list is that items in the list do not need to be of the same type[18].
2. **Dictionary** – is the most used data type. Each key is separated from its value by a colon (:), the items are separated by commas and the whole thing is enclosed in curly braces. An empty dictionary is written just with two curly braces for example: {}. Keys are unique within a dictionary where values do not have to be. Values can be of any data type, but keys must be of an immutable data type such as strings, numbers or tuples[19].
3. **Set** – this data type is a collection type. It has been with python since version 2.4. A set contains an unordered collection of unique and immutable objects. The set data type is a python implementation of the sets as they are known from mathematics which is why sets unlike lists or tuples cannot have multiple occurrences of the same element[20].
4. **Tuple** – is a sequence of immutable Python objects. Tuples, just like lists are sequences. Tuples differ from lists in a way that they cannot be changed and are parentheses unlike lists which use square brackets[21].

These four data types can be very useful for extracting required information from tweets and analysing it to determine whether tweets are positive, negative or neutral.

Over all python is a very simple and minimalistic language. Reading good python code is almost like reading very strict English. Python has a pseudo-code nature which allows you to concentrate on the solution to the problem rather than the language itself. With extraordinary simplistic syntax python makes it easy to get started with for programmers who have little to no experience with it. As python is a high-level language, when writing code – you never need to worry about the low-level details such as managing the memory used by your programs, because python manages it for you.

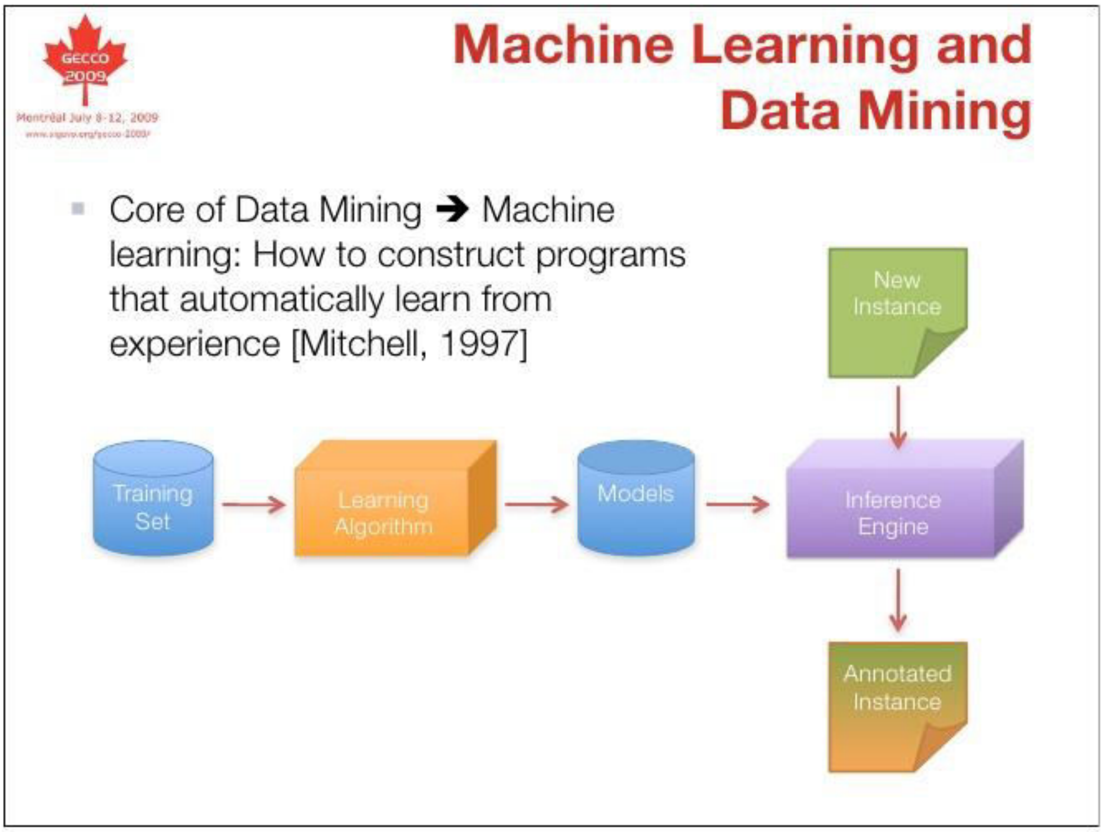
Since python has a portable nature, a programmer can produce a software product which will run on many different platforms unless system-dependent features are included.

Python is extensible. It allows you to code parts of your program in other languages such as C++ or C which then can be used from your python program.

Due to the huge python database, programmers are able to do various things involving regular expressions, documentation, generation, unit testing, threading, databases, web browsers, Common Gateway Interface (CGI), File Transfer Protocol (FTP), email, Extensible Markup Language(XML), Extensible Markup Language Remote Procedure Call protocol (XML-RPC), HyperText Markup Language (HTML), Waveform Audio File Format (WAV) files, cryptography, Graphical User Interface (GUI), Tool Kit (TK) and other system-dependent stuff. All of these features are available everywhere where python is installed.

3.4 Machine learning algorithms

Machine learning is a subfield of computer science which was developed from the investigation of pattern recognition and computational learning hypothesis in artificial intelligence[22]. Machine learning investigates the study and development of algorithms that can gain from information and make predictions on it. Such algorithms act by building a model from example inputs in order to make data-driven predictions and decisions rather than following strictly static program instructions as can be seen in figure 2 below which shows how machine learning operates.



Machine Learning and Data Mining

- Core of Data Mining ➔ Machine learning: How to construct programs that automatically learn from experience [Mitchell, 1997]

Training Set → Learning Algorithm → Models → Inference Engine

New Instance → Inference Engine → Annotated Instance

Figure2. Large scale data mining using genetics based machine learning[23].

3.5 Overview

Arthur Samuel characterised machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed" in 1959.
Tom M. Mitchell gave a generally cited, more formal definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its

performance at tasks in T, as measured by P, improves with experience E". This definition is prominent for its characterizing machine learning in on very basic level operational rather subjective terms. Taking after Alan Turing's proposition which expresses that the inquiry "Can machine think?" ought to be supplanted with the inquiry "Can machines do what we can do?"[24].

Machine learning has many tasks, however they are typically classified into three broad categories, depending on the nature of the learning "signals" or "feedback" available to the learning system. These three categories are:

- **Supervised learning** – this task uses a training data which consists of training examples. Each example is a pair consisting of an input object mostly a vector and a desired output value called supervisory signal. A supervisory learning algorithm analyses the training data and produces an inferred function which can be used for mapping new examples. In order to solve supervisory learning problem one has to perform a series of steps such as[25][26][27]:
    - Determine the type of training examples.
    - Gather a training set.
    - Determine the input feature representation of the learned function.
    - Determine the structure of the learned function and corresponding learning algorithms.
    - Complete a design.
    - Evaluate the accuracy of the learned function.

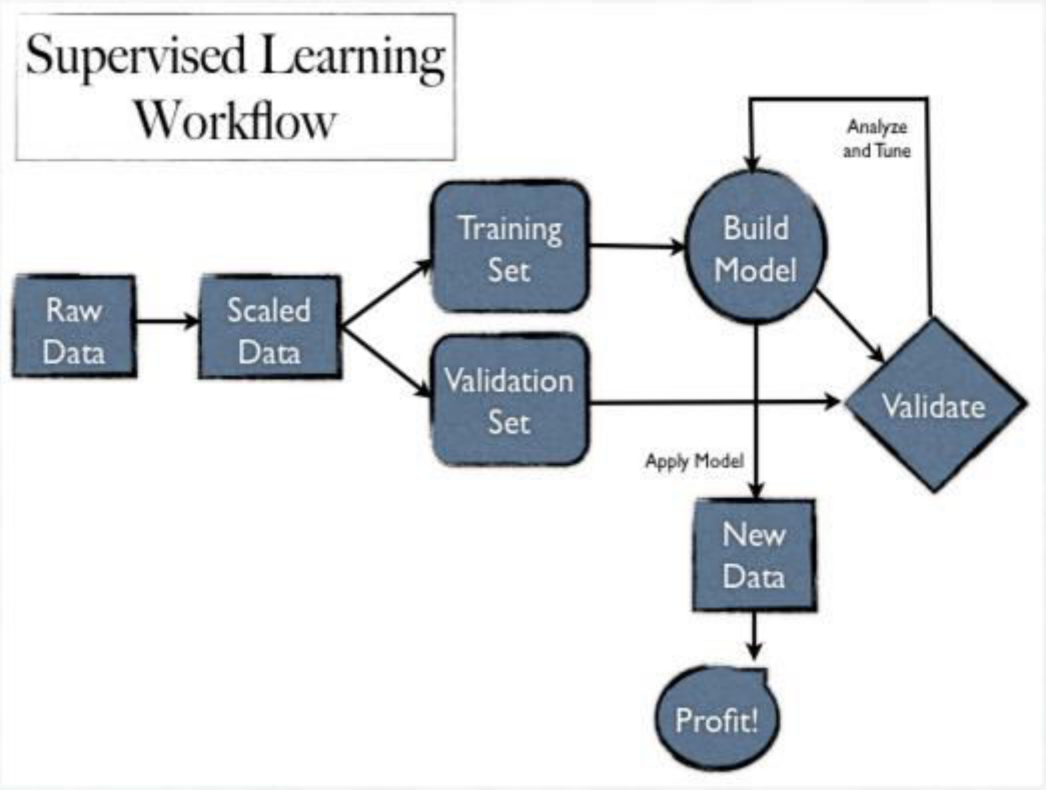


Figure3. Supervised learning workflow[25].

- **Unsupervised learning** – there are no labelled examples given to the algorithm and as such there are no error or reward signal to evaluate a potential solution. Many methods employed in unsupervised learning are based on data mining methods used to process data. The problem of unsupervised learning is that of trying to find hidden structure in unlabelled data[25][26][27].

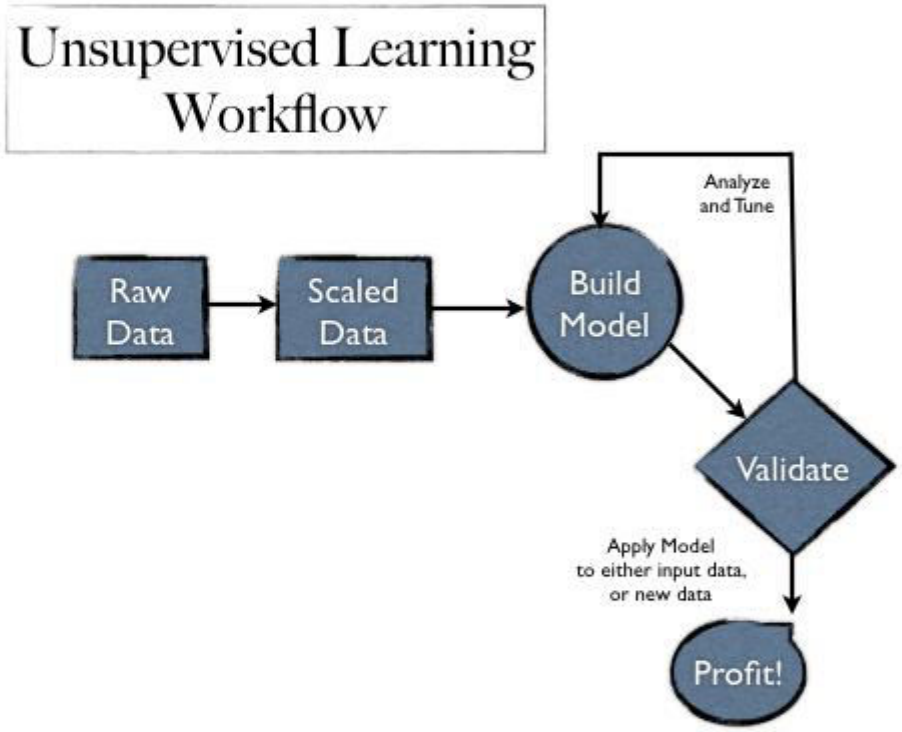

Unsupervised Learning Workflow
Raw Data
Scaled Data
Build Model
Analyze and Tune
Validate
Apply Model to either input data, or new data
Profit!

Figure4. Unsupervised learning workflow[25].

- **Reinforcement learning** – a computer program must perform a certain goal in dynamic environment that it interacts with. It must perform this goal without a teacher explicitly telling it whether it has come close or not. An example of this can be learning to play a game by playing against an opponent. As you can see below in figure 5, reinforcement learning allows the machine or software to learn its behaviour on feedback from the environment[25][26][27].

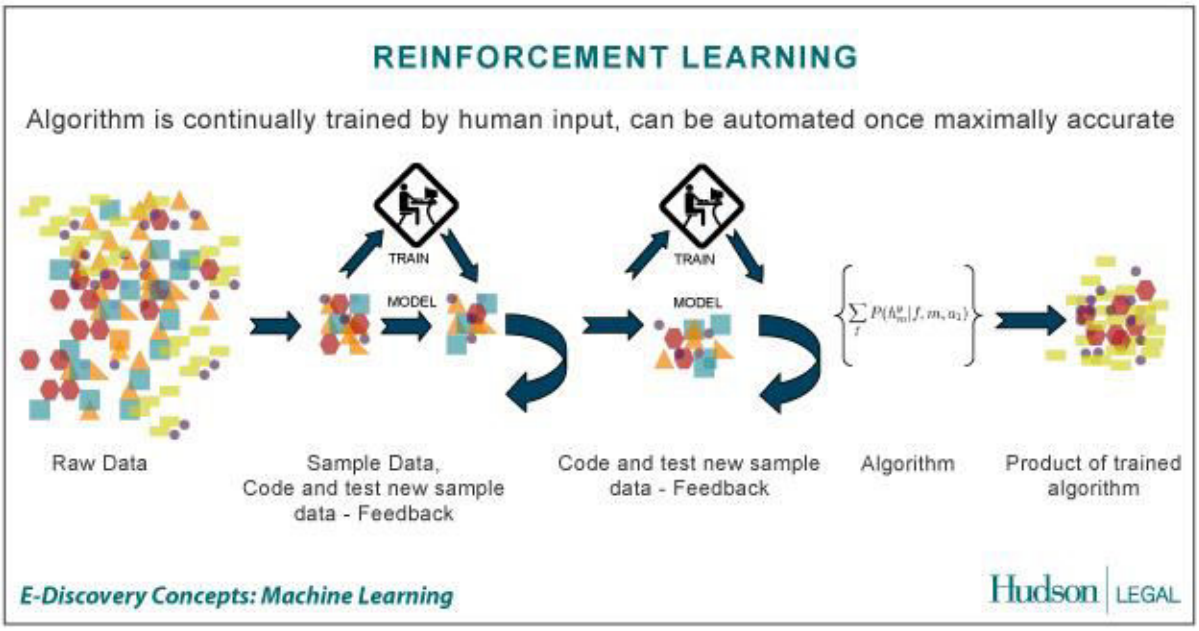

REINFORCEMENT LEARNING
Algorithm is continually trained by human input, can be automated once maximally accurate
TRAIN
MODEL
TRAIN
MODEL
Raw Data
Sample Data, Code and test new sample data - Feedback
Code and test new sample data - Feedback
Algorithm
Product of trained algorithm
E-Discovery Concepts: Machine Learning
Hudson LEGAL

Figure5. Reinforcement learning[25].

3.6 Approaches

There are many approaches for machine learning algorithms. I will list briefly some of these approaches:

- **Decision tree learning** – this approach utilises a choice tree as a prescient model, which maps perceptions about an item to conclusions about the item's target value[28][29].
- **Association rule learning** – is a method for finding intriguing relations between variables in expansive databases[29].
- **Artificial neural networks** – this learning algorithm, otherwise called "neural network", is a learning algorithm that is motivated by the structure and functional aspects of biological neural networks. All computations are organised as far as an interconnected group of artificial neurons, handling data utilising a connectionist way to deal with the computation. Modernised neuron networks are non-linear factual information displaying tools. These neural networks are utilised to model entangled connections between inputs and outputs, look for data patterns or capture factual structure in an obscure joint likelihood distribution between observed variables[29].
- **Support vector machines** – are administered learning models with related learning algorithms. These algorithms examine data and perceive patterns utilised for classification and regression analysis. Each given training example is marked for belonging to one of two categories. The support vector machines training algorithm build a model which assigns new examples into one of the two categories, making it a non-probabilistic binary linear classifier. All examples are mapped so that they fall under one category or the other which makes as wide as possible gap between the two categories[29].

These are but a few of machine learning algorithms, others would be: **Clustering**, **Bayesian networks**, **Reinforcement learning**, **Representation learning**, **Similarity and metric learning**, **Sparse dictionary learning** and **Genetic algorithms**[29].

3.7 Alternatives

There are many alternative ways to implement sentiment analysis as it does not have limitations to a specific language or approach. Sentiment analysis can be implemented using many different languages such as C#, C++, Java and others. I will discuss some of these languages as the alternatives and compare them to python:

- **C#** - Is an easy to learn and use language. It is a multi-paradigm programming language which accompanies strong typing, imperative, declarative, functional, generic, object-oriented and component-oriented programming disciplines. C# was developed by Microsoft, and is designed for Common Language Infrastructure. This programming language is intended to be a simple, modern, general-purpose and object oriented[30].
    - **History** – In July 1999, Anders Hejlsberg formed a team to develop a new programming language. The language developed was C#. Its intended name was 'Cool' but was later changed to C# for trademark reasons. In July 2000 C# was published[31]. When C# first came out, a lot of people stated that it was very similar to Java as a result of that – C# was called a Java clone. Despite many people stating that C# was a clone of Java, Anders Hejlsberg said that C# is not a clone of Java and in fact is much closer to C++ in its design. Upon C # 2.0 releases both languages have diverged from each other to the point where they were not similar anymore[32].
    - **C# and Python** – Both languages are easy to learn and simple to use. Despite C# and Python having great documentation, very good reliability and both being easy to

    use, A lot of programmers prefer Python over C# due to the fact that Python focuses on allowing programmers to focus on the problem rather than having to worry about the language itself which makes Python a better option for problem solving than C# and that is why for this project I think Python is more suitable[33].

- **Java** – is one of the most popular programming languages used today, mainly because of client-server web applications, with a reported nine million developers. Java is a general-purpose programming language which is concurrent, class-based, object-oriented and specifically designed to have as few implementation dependencies as possible. This programming language was designed in mind of allowing programmers to write code once and run it anywhere, which means that compiled java code can run on any java supported platform[34].
    - **History** – Java project was initiated in June 1991 by James Gosling, Mike Sheridan and Patrick Naughton. It was originally designed for interactive television, but was too advanced for the digital cable television industry at the time. Initially Java language was called Oak, latter it was renamed Green and finally renamed to Java. This programming language was designed with C/C++ style syntax in mind so that system and application programmers would find it familiar. First public implementation of Java was released in 1995 by Sun Microsystems, it was called Java 1.0. This implementation promised to “Write Once, Run Anywhere”, on popular platforms while providing no cost run-times. New version of Java called Java 2 was released in December 1998 – 1999 with multiple configurations built for different platform types. In 2006 Java 2 versions J2EE, J2ME and J2SE were renamed to Java EE, Java ME and Java SE respectively[35].
    - **Java and Python** – Java programs execute faster than python programs, but python programs are three to five times shorter and easier to code than java programs thanks to python’s dynamic typing. In java every variable needs to be explicitly declared because java is statically typed language, where in python a programmer wastes no time in declaring variables. This takes python programs longer to execute because python’s run time works harder than java’s. It needs to evaluate objects to find their types. Even though python programs take slightly longer to execute, in my opinion tackling sentiment analysis using python will be faster and easier as I will not have to worry about the language itself and will be able to focus on the problem. This is why python is a better choice for this particular project[36].
- **C++** – Is a language that influenced other languages such as: Java and C#. It is a general-purpose programming language. This language has imperative, object-oriented and generic programming features. It also provides facilities for low-level memory manipulation. C++ programs tend to execute and run fast compared to other high level programming languages, because it was designed with a bias toward system programming. C++ design highlights are performance, efficiency and flexibility of use[37].
    - **History** – C++ was influenced by many languages including: C, Simula, ALGOL 68, Ada, CUL and ML. At first C++ was called just C with Classes because the class, derive class, strong typing, inlining and default argument features were added to C, in 1983 it was renamed to C++. The first edition of C++ programming language was release in 1985 and became the definitive reference for the language as there was not an official standard yet. C++ 2.0 was released in 1989 followed by The C++ Programming Language being released in 1991. Multiple improvements and new features come with each release up to the present version. Various new additions are planned for 2017 version[37].
    - **C++ and Python** – what has been told in the comparison of java and python can be applied here and even more so. Python code is five to ten times shorter than C++ code and it is said that what a python programmer can complete in two months, two

C++ programmers cannot complete in a year. That being said – C++ code of course executes and runs faster than python's code. Taking everything into account I have easily determined that despite python's programs executing slightly slower, the fact that programming in python is so much easier, out weights C++ speed. C++ code can be tedious and difficult to understand where python's code will be easy to code and understand making python the better choice for this project[36].

3.8 Technology pick

Different approaches exist for implementing sentiment analysis, one of them being simple key word spotting. Programmer defines a list of key words with positive, negative or neutral ratings that the program will look work in tweets. Once the words are found and picked out – tweet rating is calculated determining whether it is positive, negative or neutral. This approach to sentiment analysis is very bad and naïve, but very easy for simple sentiment analysis programs.

For this particular project as I have discussed above I have chosen python as the programming language to be used to implement sentiment analysis for Twitter. As for the approach – for the sakes of actually coding a working sentiment analysis program which analyses Twitter tweets I will use key word spotting. When the program will be fully functional and sentiment analysis will be implemented using this approach – I will look into implementing machine learning algorithm approach for better analysis of tweets. The end product will be built using python and sentiment analysis will be implemented using machine learning algorithm approach.

4. Database

This project will require a data base to store tweets. There are numerous databases on World Wide Web (WWW). I will pick few most popular databases and discuss them. In the end I will compare all databases I have picked and will choose the best option for this project.

4.1 Microsoft SQL Server Express

This is a free version to download and use of Microsoft's SQL Server retaliation database management system. This database is targeted for smaller scale applications. Even though this version is free, it still provides many features that of the paid version. Features that are provided by the free express version are as follows[38]:

- Database size – paid version provides a maximum of 524PB (petabytes) where's the free version provides only 10GB (gigabytes) of free storage. This is very small amount compared to the paid version, but more than enough for this project[38][39].
- Hardware usage is allowed, but it is limited to a single physical CPU and 1GB of Random Access Memory (RAM). Despite these limitations, the database is still very good as it allows the use of multiple cores[38][39].
- In previous versions of SQL Express, a concurrent workload- governor was included to limit the performance of the database if it receives more work than is typical of a small number of users, however this is not the case in current versions[38][39].
- Sever GUI tools are included for the database management. Such as:
    - SQL Server Management Studio Express – software is used to configure, manage and administer all components within Microsoft SQL Server. This software includes script editors and graphical tools which work with the features and objects on the server[40].

- SQL Server Configuration Manager – is a tool to deal with the services connected with SQL Server, to arrange the network protocols utilised by SQL Server, and to deal with the network availability setup from SQL Server client computers[41].
- SQL Server Surface Area Configuration tool - is a security measure that includes ceasing or incapacitating unused segments. Surface territory diminishment enhances security by giving fewer boulevards to potential assaults on a framework[42].
- SQL Server Business Intelligence Development Studio – is an essential environment for creating business solutions that incorporate Analysis Services, Integration Services and Reporting Services projects[43].

While this free version includes these features, it also does not include some useful features which are included in the full version. Such features would be:

- SQL Server Agent service – accommodates features permitting the planning of periodic activities on Microsoft SQL Server 2000, or the warning to system administrator of issues that have happened with the server[44].
- Bigger storage capacity as was mentioned before[39].
- Faster database overall performance[39].

4.2 MySQL

MySql provides a best-of-all world’s scenario in a lot of ways: It is supported by many platforms, appreciates low total cost of ownership (TCO) and is stable. MySQL has excellent documentation. MySQL AB has an exhaustive web site containing reference material and additionally a connection to mailing-list archives. High quality support is also provided by MySQL AB which includes a service that permits MySQL developers to sign into your server to correct problems and proactively help with optimization. MySQL offers stability, support and low cost which gains it relational database management system (RDBMS) market share[45].

Over all, this database is most popular in the world, because of its proven performance, reliability and ease-of-use.

This database is good at many things and provides many features for the user, although the things that it is best at would be as follows[46]:

- **Web applications** – normally include numerous reads and few writes. Since MySQL is quick, it meets the requests of internet speed. In many programmers’ experience MySQL has proven to outperform other RDBMS products in web applications.
- **Enterprise-level applications** – MySQL provides support directly through the parent company which is MySQL AB. Feature wise MySQL provides just about everything that would be required by an enterprise-level application.
- **Open-source support** – MySQL is open source which makes it available for everyone to download it and extend the code to meet his or her needs.
- **Low overhead** – MySQL does not require a lot of computer resources and so it can be easily run even on Intel Premium-class computer which normally has 32 megabytes (MB) of RAM or less. Although it is not recommended to run an enterprise-level MySQL implementation on such a system, because it would require quite much more computer resources.
- **Available large table size** – unlike other RDBMS products, MySQL provides large table sizes for free, though some file-size limitations of the host operating system can be encountered. It has been tested that some architectures can support up to eight terabytes (TB) per table.
- **Stability** – MySQL’s software is always in development. Some features are newer than others, which can lead to them being less stable. Over all though, MySQL is very reliable and stable database management system as can be seen in figure 6 below.

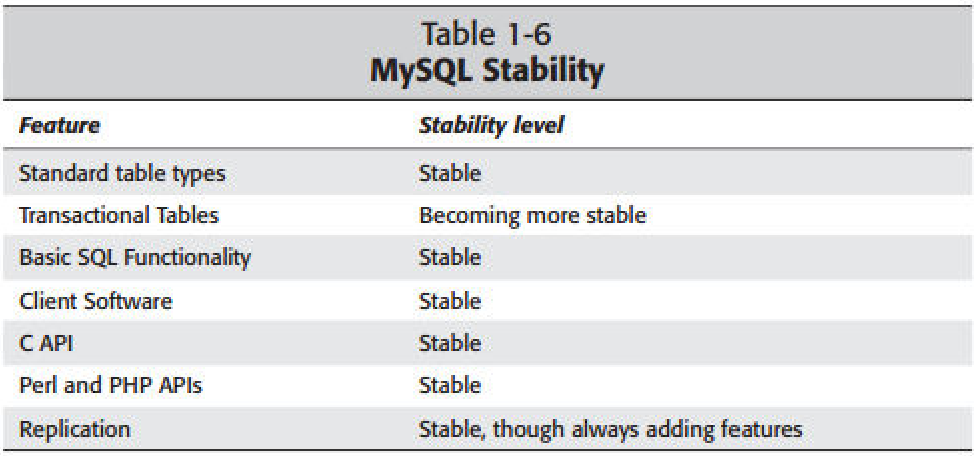


Table 1-6
MySQL Stability

Feature	*Stability level*
Standard table types	Stable
Transactional Tables	Becoming more stable
Basic SQL Functionality	Stable
Client Software	Stable
C API	Stable
Perl and PHP APIs	Stable
Replication	Stable, though always adding features

Figure6. Stability of MySQL database management system[46].

MySQL beats even some of its own commercial counterparts when it comes to performance, scalability and stability. It can perform just as good or even better as its competitors. Few popular features of MySQL can be seen compared to other RDBMS products in figure 7 below.

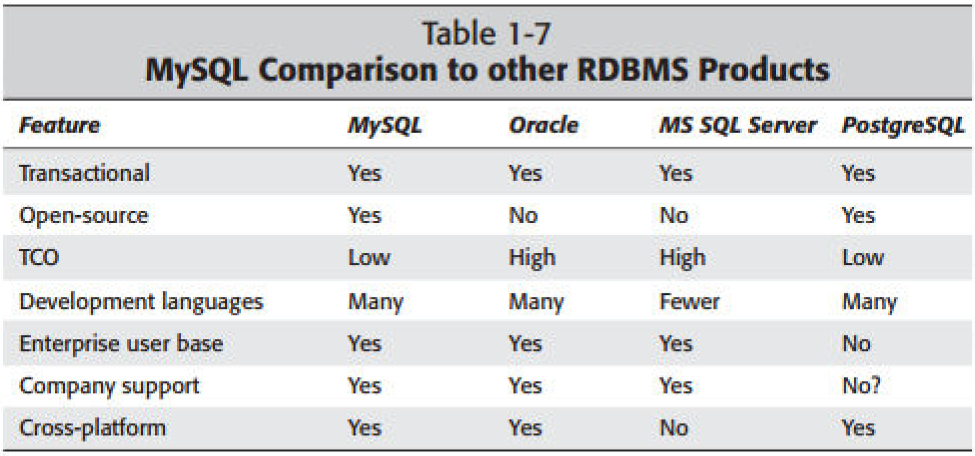


Table 1-7
MySQL Comparison to other RDBMS Products

Feature	*MySQL*	*Oracle*	*MS SQL Server*	*PostgreSQL*
Transactional	Yes	Yes	Yes	Yes
Open-source	Yes	No	No	Yes
TCO	Low	High	High	Low
Development languages	Many	Many	Fewer	Many
Enterprise user base	Yes	Yes	Yes	No
Company support	Yes	Yes	Yes	No?
Cross-platform	Yes	Yes	No	Yes

Figure7. Comparison of MySQL popular features with other RDBMS products[46].

Even with all features that MySQL provides, there are things that it cannot do yet[46]:

- **Foreign keys** – are values that relate to the Primary keys in another table. This feature is popular and is used often in Oracle and other RDBMS products. MySQL has already started support for foreign keys in its version 4.0 and will enhance this feature in future. This feature is to be implemented as of yet.
- **Inherited tables -** will not be included as they are not even planed for any MySQL versions as of yet.

We looked in features that are and are not supported by MySQL and although some popular features such as foreign keys are not supported, this database is number one pick for many programmers and users world-wide. MySQL provides stability, performance, low resource demand, huge table size, reliability and ease of use making it so far the number one pick for this project.

4.3 MariaDB

MariaDB is a double for MySQL. It was created by several former core developers of MySQL. These developers left the company being unsatisfied by poor quality and rate of improvement after it was acquired by Sun Microsystems in 2008 which in turn was later acquired by Oracle Corporation in 2009. MariaDB was developed as an enhanced, drop-in, binary compatible replacement for MySQL[47].

The developers took freely available MySQL code and enhanced it. Because of that MariaDB has numerous similarities to MySQL. When changes are made to MySQL, they are also made to MariaDB by developers reviewing and incorporating the same changes.
Some similarities between the two database management systems would be[47]:

- **Same file names** – although when installing MariaDB, the downloaded package's name is different from that of the MySQL, the names of installed files and binaries are identical to that of MySQL.
- **Language** – the SQL language for bot databases is identical and so are the configuration files with the exception of few new MariaDB parts that are easy to learn.
- **Client programs** – from things like ports, sockets and client APIs, are identical between the two databases. Every language such as Java, C, PHP, .Net, Perl, Python and Ruby that can talk to MySQL database, can also talk to MariaDB database because the work of MySQL connectors is unchanged in MariaDB.
- **Databases** – each database management system can open databases created by the other.
- **Learning** – any experienced MySQL database user or database administrator (DBA) will have no problem transferring to MariaDB database management system. The differences are really minor and anyone with experience in MySQL will be able to learn and take advantage of those differences very quickly with little effort.

The similarities between the two database management systems are endless, but there are differences also, because otherwise there would be no use to have two identical database management systems. Some essential differences would be as follows:

- **Ease of Use** – MariaDB developers found and implemented few ways to make the lives of this database management system users easier. Some of these ways would be[47]:
  - **User Statistics** – this feature provides the statistics of INDEX and TABLE. It ads few new information schema tables few new FLUSH and SHOW commands. Using these commands the server activity can be understood better and sources of your database's loads can be identified.
  - **ALTER TABLE and LOAD DATA INFILE commands** – improves long running. These commands show you how much progress you have made by receiving a progress message from the server. This can be very useful, especially since you do not want to cancel a long-running ALTER TABLE command which is almost complete.
  - **Microsecond support** – provides more accurate TIME, DATETIME and TIMESTAM datatypes precision.
  - **NoSQL-style features** – would include:
    - **HandlerSocket** – using this feature a fast and direct access is provided to InnoDB tables by skipping the SQL layer.
    - **Dynamic Columns** – using this feature, users are provided with the ability to have a different set of virtual columns for each row in a table.
  - **Subqueries** – unlike in MySQL, MariaDB provides actually usable and useful subqueries.
- **Performance** – even though MySQL has great performance out of the box, it does not mean that it cannot be improved and that is one of the major goals of the MariaDB developers[47].
  - **The optimizer** – this is the engine which sits at the core of both database management systems. It takes the entered SQL commands and turns them into

    instructions for the database. MariaDB has significant improvements in this area, especially on complex workload.
  - **Replication** – another area of focus to which MariaDB developers have brought improvements, one of them would be:
    - **Group commit for the binary log** – this creates numerous setups which use replication and have many updates more than two times faster.
  - **Table Elimination** – sometimes it is possible to resolve a query without accessing some of the tables it refers to. This makes the query complete faster since it does not need to access as many tables.
- **Better testing** – a core component to writing better code is testing. MariaDB developers have made huge improvements to the testing infrastructure of MariaDB. Some of these improvements would be[47]:
  - **More tests** – numerous new tests were added to MariaDB test suit.
  - **Test suit bug fixes** – the test suit itself is tested by MariaDB developers just like any other parts of this database management system.
  - **Better feature testing** – are achieved by testing builds with different configuration options on multiple operating systems and processor combinations.
  - **Removal of invalid tests**.
- **Fewer Bugs and Warnings** – numerous efforts are made to fix as many bugs as possible in all MariaDB releases. This is achieved with the help of enhanced MariaDB testing infrastructure and deep knowledge and experience of MariaDB developers. Related issue is the compiler warnings. According to MariaDB developers, these are almost as bad as bugs, which is why a lot of effort is put in to fixing as many of them as possible[47].

These are just three of numerous database management systems world-wide. Out of these three which I have discussed in detail, MySQL is a great database management system. It is reliable, fast and full of features which can be used by anyone for free all around the world. This database management system is a number one pick for many users.

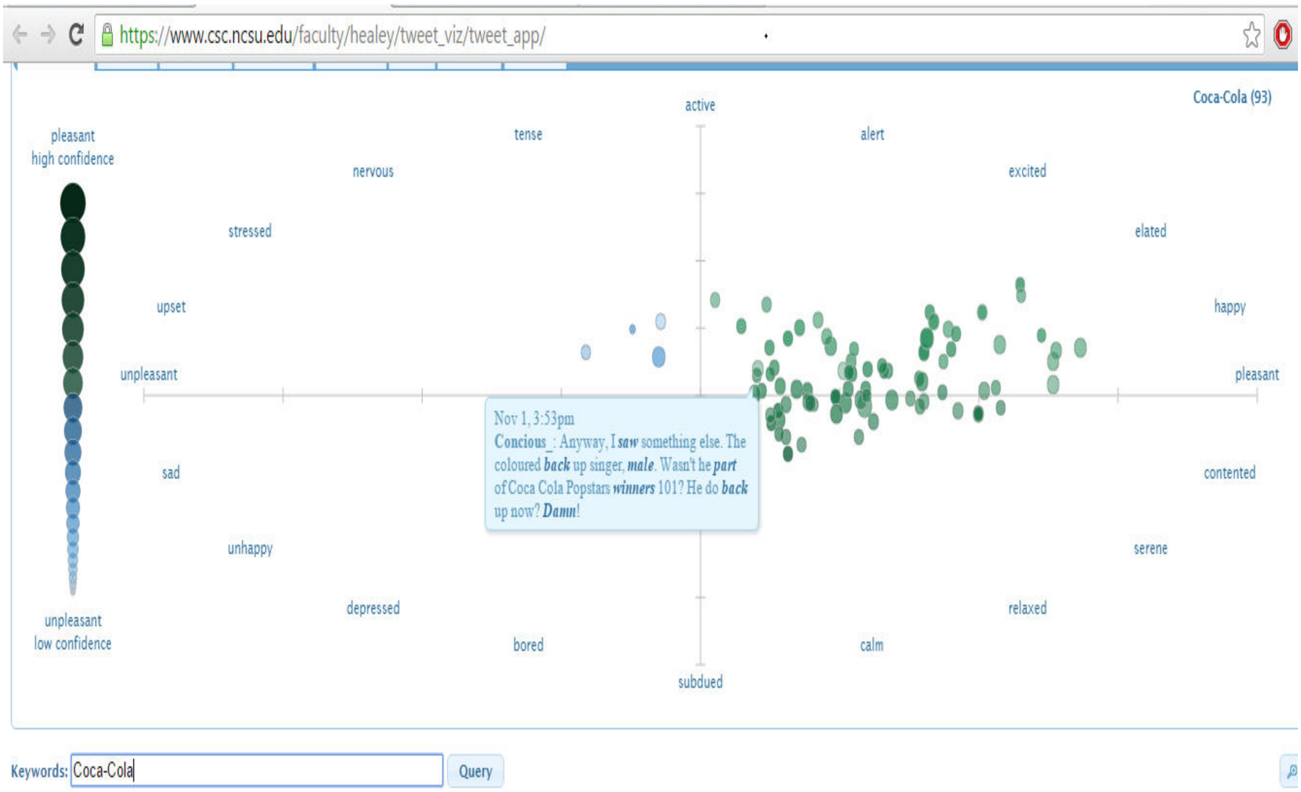
With that in mind I decided that MariaDB is the best pick overall. With its similarity to MySQL making it as easy and even easier to use, improved speed, stability, reliability and overall performance makes it a future of MySQL database and the database that will be used for this product.

5. Related work

Sentiment analysis is the ever growing area of interest in past years. More and more companies require some sort of sentiment analysis implementations to review their products. There are a lot of sentiment analysis applications for reviewing tweets already. In this section I will list and briefly discuss some of these applications.

5.1 Tweet Sentiment Visualisation App

This is a web application which will extract tweets that are relevant to your query, analyse them and place them in a circle displaying overall result.



Nov 1, 3:53pm
Concious_: Anyway, I ***saw*** something else. The coloured ***back*** up singer, ***male***. Wasn't he ***part*** of Coca Cola Popstars ***winners*** 101? He do ***back*** up now? ***Damn***!

Keywords: Coca-Cola

Figure8. Tweet Sentiment Visualisation application[48].

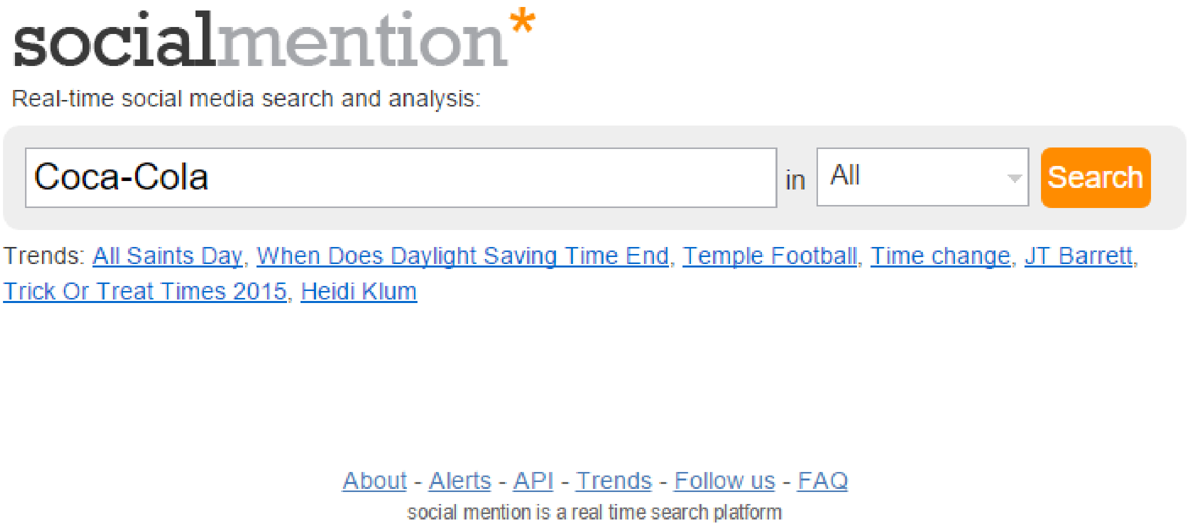
As can be seen in figure 8 above, this application rather nicely displays related tweets in a circle according to their evaluation of emotion. Tweets are represented like small circles which can be clicked on and the contents of the actual tweet will be displayed. The evaluation scheme looks rather complex and accurate which makes this application quite reliable.

With this kind of representation of tweets and the overall result, while graphically appealing, might be inconvenient for quick analysis and understanding of overall peoples' emotional response of a specific company's product.

Over all this application is well made with beautiful GUI and quite well made features such as zoom in and out for better separation of individual tweets, and extra tabs to choose from for more functionality. That said, an application which would display simple results stating whether people like or dislike the product might be more efficient.

5.2 Socialmention

Is a web application which searches for your query and displays results from different sources of your choice.



Real-time social media search and analysis:

Coca-Cola

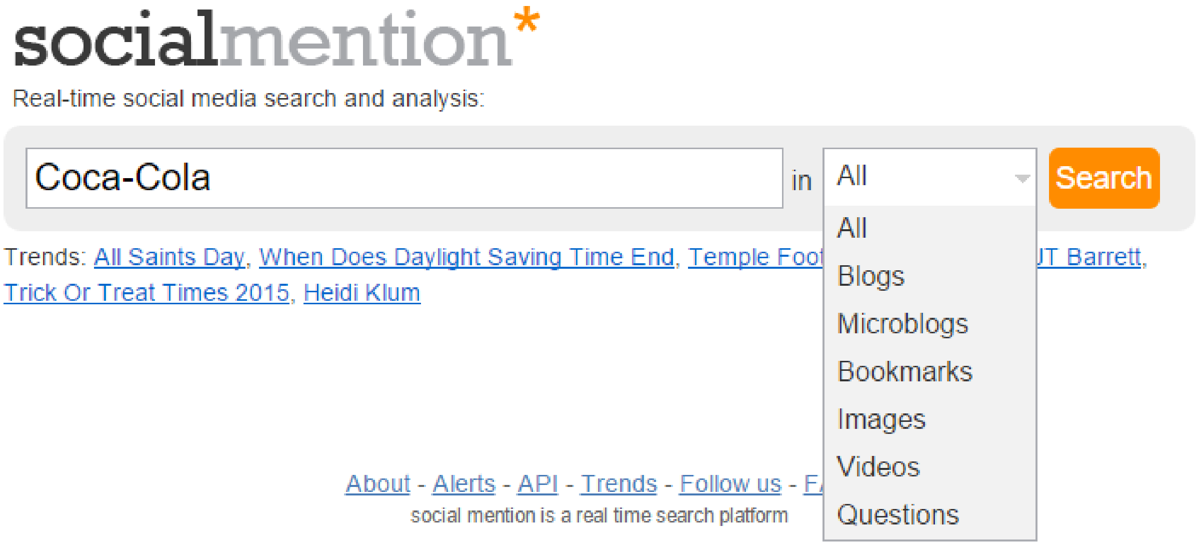
Trends: All Saints Day, When Does Daylight Saving Time End, Temple Football, Time change, JT Barrett, Trick Or Treat Times 2015, Heidi Klum

About - Alerts - API - Trends - Follow us - FAQ

social mention is a real time search platform

Figure9. Search page of socialmention web application[49].

As can be seen in figure 9 above the search page is quite simple and efficient. User is provided with a text field for his query. On the right user can select from which source relevant posts will be extracted and analysed. This can be seen in figure 10 below.



Real-time social media search and analysis:

Coca-Cola

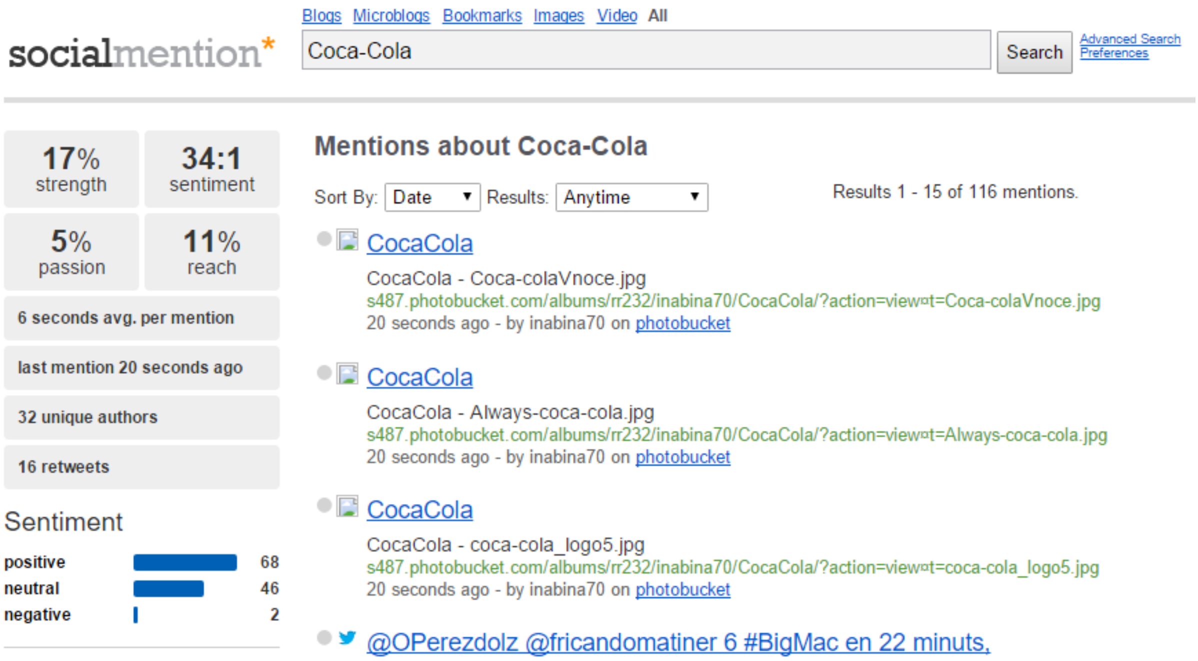
Trends: All Saints Day, When Does Daylight Saving Time End, Temple Foot JT Barrett, Trick Or Treat Times 2015, Heidi Klum

About - Alerts - API - Trends - Follow us - F

social mention is a real time search platform

Figure10. Socialmention search options[49].

Once user hits the search button, socialmention will look through selected sources and extract relevant posts. Those posts then are analysed and overall results are displayed on the left, while extracted posts are displayed on the right.



17% strength
34:1 sentiment
5% passion
11% reach
6 seconds avg. per mention
last mention 20 seconds ago
32 unique authors
16 retweets

Sentiment

positive	68
neutral	46
negative	2

Mentions about Coca-Cola

Sort By: Date Results: Anytime Results 1 - 15 of 116 mentions.

CocaCola
CocaCola - Coca-colaVnoce.jpg
s487.photobucket.com/albums/rr232/inabina70/CocaCola/?action=view¤t=Coca-colaVnoce.jpg
20 seconds ago - by inabina70 on photobucket

CocaCola
CocaCola - Always-coca-cola.jpg
s487.photobucket.com/albums/rr232/inabina70/CocaCola/?action=view¤t=Always-coca-cola.jpg
20 seconds ago - by inabina70 on photobucket

CocaCola
CocaCola - coca-cola_logo5.jpg
s487.photobucket.com/albums/rr232/inabina70/CocaCola/?action=view¤t=coca-cola_logo5.jpg
20 seconds ago - by inabina70 on photobucket

@OPerezdolz @fricandomatiner 6 #BigMac en 22 minuts,

Figure11. Socialmention results[49].

As can be seen in figure 11 above, this web application displays required information rather conveniently, making it easy for users to navigate through and understand it.

Over all this application is well built, simple and easy to use. While this application has quite more features than what is planned for this project, it is something like this that the project will look like.

5.3 Overall

These are but two of many applications that are similar to what this project is aiming for. Some of them provide simple GUI with efficient outputs of results while others use more advanced visuals making the application look attractive but not as efficient.

6. Conclusion

We have looked at the background of sentiment analysis, how it has started to gain interest and failed, gut recently revived and is an ever growing area in interest and importance for companies. We have looked at different ways sentiment analysis can be implemented and have discussed some of them in detail. We looked in a few different popular programming languages which potentially can be used to implement sentiment analysis. We have realised that sentiment analysis is not bound to one or few programming languages, but can be implemented using any.

I have decided that python is the best programming language to go about implementing sentiment analysis for this project as python is a high level languages which makes coding as easy and short as possible.

I have discussed many different ways the sentiment analysis implementation can be approached and decided to use the simplest but also the most naïve way called keyword spotting. This method picks out keywords from tweets and measures the sentiment of tweets by weights associated to those keywords. This approach is very bad when dealing with sarcasm, but is very easy to implement and so for the sakes of simplicity and getting the project going I chose to use this approach at the beginning. Once the basics will be implemented, I will try and implement machine learning algorithms for better evaluation of tweets.

We have looked at three databases from which I have picked MariaDB database management system to be used for this project as it provides best reliability, ease of use and overall performance compared to the other database management systems we have discussed.

We have looked at similar applications and found that even if they are not quite what this project is aiming for, those applications provide a lot of functionality and visualisation which could provide ideas for this project.

Overall this application will be a web application which will retrieve tweets, store them on MariaDB database and evaluate them according to search queries. The application will be built using python and keyword spotting approach at first, latter if everything works – machine learning algorithm will be utilised to improve over all evaluation. The GUI will be as simple as possible to make it easy to use and understand. Evaluated tweets will be displayed in the same page as the overall result so users can re-evaluate manually if needed.

References

1. Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
2. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
3. Pang, Bo, and Lillian Lee. "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005.
4. Snyder, Benjamin, and Regina Barzilay. "Multiple Aspect Ranking Using the Good Grief Algorithm." HLT-NAACL. 2007.
5. Science Direct. (2014). Sentiment analysis process on product reviews. [online], available: http://www.sciencedirect.com/science/article/pii/S2090447914000550 [accessed 20 October, 2015].
6. Pang, Bo, and Lillian Lee. "4.1. 2 Subjectivity Detection and Opinion Identification." Opinion Mining and Sentiment Analysis. Now Publishers Inc.(2008).
7. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews."Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
8. Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis."IEEE Intelligent Systems 2 (2013): 15-21.
9. Cambria, Erik, Catherine Havasi, and Amir Hussain. "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis." FLAIRS conference. 2012.
10. Maria Ogneva. (2010). How Companies Can Use Sentiment Analysis to Improve Their Business. Mashable.19 April, [online], available: http://mashable.com/2010/04/19/sentiment-analysis/#3gKvnSLBFiqc [accessed 22 October 2015].

11. Jamie Condliffe. Flaming drives online social networks, NewScientist, 2010-12-07, available: https://www.newscientist.com/article/dn19821-flaming-drives-online-social-networks [accessed 22 October 2015].
12. TIOBE Software Index (2015). TIOBE Programming Community Index Python, [online], available: http://www.tiobe.com/index.php/paperinfo/tpci/Python.html [accessed 22 October 2015].
13. Tecosystems (2015). The RedMonk Programming Language Rankings: June 2015, [online], available: http://redmonk.com/sogrady/2015/07/01/language-rankings-6-15/ [accessed 22 October 2015].
14. Awaroop C.H. (2013). A Byte of Python. Kindle Edition. Ebshelf, Inc. [online] available: http://files.swaroopch.com/python/byte_of_python.pdf [accessed 23 October 2015].
15. Bill Venners (2003). The Making of Python. Artima Developer, [online], 13 January, available: http://www.artima.com/intv/pythonP.html [accessed 23 October 2015].
16. Guido van Rossum (2009). A Brief Timeline of Python. The History of Python, Google, [online], 20 January, available: http://python-history.blogspot.ie/2009/01/brief-timeline-of-python.html [accessed 23 October 2015].
17. Python Software Foundation (2015). Python 3.0 Release, [online], available: https://www.python.org/download/releases/3.0/ [accessed 23 October 2015].
18. Tutorialspoint (2015). Python Lists, [online], available: http://www.tutorialspoint.com/python/python_lists.htm [accessed 24 October 2015].
19. Tutorialspoint (2015). Python Dicttionary, [online], available: http://www.tutorialspoint.com//python/python_dictionary.htm [accessed 24 October 2015].
20. Python Course (2015). Sets and Frozensets, [online], available: http://www.python-course.eu/sets_frozensets.php [accessed 24 October 2015].
21. Tutorialspoint (2015). Python Tuples, [online], available: http://www.tutorialspoint.com//python/python_tuples.htm [accessed 24 October 2015].
22. Encyclopedia Britannica (2015). Machine learning Artificial intelligence, [online], available: http://www.britannica.com/technology/machine-learning [accessed 25 October 2015].
23. Jaume Bacardit, Xavier Llora. (2009). Larger Scale Data Mining using Genetics-Based Machine Learning. [online], available: http://www.slideshare.net/xllora/large-scale-data-mining-using-geneticsbased-machine-learning [accessed 25 October 2015].
24. Harnad, Stevan (2008), The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence, in Epstein, Robert; Peters, Grace, The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer, Kluwer
25. Ron Kohavi; Foster Provost (1998). Glossary of terms, Machine Learning, [online], 8 May 2014, available: https://translate.google.com/translate?hl=en&sl=auto&tl=en&u=http%3A%2F%2Fyemimamikha.blog.binusian.org%2F2014%2F05%2F12%2Ftugas-t0264-intelegensia-semu-gslc-2-8-mei-2014%2F&sandbox=1 [accessed 25 October 2015].
26. Scikits learn (2015). Machine Learning 101: General Concepts, [online], available: http://www.astroml.org/sklearn_tutorial/general_concepts.html [accessed 25 October 2015].
27. Ajay Singh. (2014). Data Science with Hadoop. slideshare, [online], 19 December, available: http://www.slideshare.net/hortonworks/data-science-workshop [accessed 25 October 2015].
28. Pedro Domingos. Decision Trees [video online], available: https://class.coursera.org/machlearning-001/lecture [accessed 27 October 2015].
29. Wikipedia The Free Encyclopedia. (2015). Machine learning, available: https://en.wikipedia.org/wiki/Machine_learning#cite_note-2 [accessed 27 October 2015].

30. Microsoft (2015). Introduction, [online], available: https://msdn.microsoft.com/en-us/library/aa645597(v=vs.71).aspx [accessed 30 October 2015].
31. Naomi Hamilton. (2008). The A-Z of Programming Languages: C#. Computerworld, [online], 01 October, available: http://www.computerworld.com.au/article/261958/a-z_programming_languages_c_/ [accessed 30 October 2015].
32. Klaus Kreft, Angelika Langer. (2003). After Java and C# - what is next?. Artima developer, [online], 03 July, available: http://www.artima.com/weblogs/viewpost.jsp?thread=6543 [accessed 30 October 2015].
33. OnStartups (2015). Python vs. C#: Business and Technology Tradeoffs, [online], available: http://onstartups.com/tabid/3339/bid/128/Python-vs-C-Business-and-Technology-Tradeoffs.aspx [accessed 31 October 2015].
34. James Gosling, Bill Joy, Guy Steele, Gilad Bracha, Alex Buckely. (2015). The Java Language Specification. Java SE 8 Edition, Oracle America, Inc., [online] available: https://docs.oracle.com/javase/specs/jls/se8/jls8.pdf [accessed 31 October 2015].
35. Jon Byous. (2005). Java Technology: The Early Years. Sun, 20 April, available: http://web.archive.org/web/20050420081440/http://java.sun.com/features/1998/05/birthday.html [accessed 1 November 2015].
36. Python (2015). Comparing Python to Other Languages, [online], available: https://www.python.org/doc/essays/comparisons/ [accessed 1 November 2015].
37. Wikipedia The Free Encyclopedia. (2015). C++, available: https://en.wikipedia.org/wiki/C%2B%2B#cite_note-Stroustrup1-3 [accessed 1 November 2015].
38. Wikipedia The Free Encyclopedia. (2015). SQL Server Express, available: https://en.wikipedia.org/wiki/SQL_Server_Express [accessed 2 November 2015].
39. Microsoft (2015). SQL Server and the data platform, available: https://www.microsoft.com/en-us/server-cloud/products/sql-server-editions/overview.aspx [accessed 2 November 2015].
40. Microsoft (2015). Use SQL Server Management Studio, available: https://msdn.microsoft.com/en-us/library/ms174173.aspx [accessed 2 November 2015].
41. Microsoft (2015). SQL Server Configuration Manager, available: https://msdn.microsoft.com/en-us/library/ms174212.aspx [accessed 2 November 2015].
42. Microsoft (2015). Surface Area Configuration, available: https://msdn.microsoft.com/en-us/library/ms161956.aspx [accessed 2 November 2015].
43. Microsoft (2015). Introducing Business Intelligence Development Studio, available: https://msdn.microsoft.com/en-us/library/ms173767(v=sql.105).aspx [accessed 2 November 2015].
44. Microsoft (2015). SQL Server Agent, available: https://msdn.microsoft.com/en-us/library/ms189237.aspx [accessed 2 November 2015].
45. MySQL (2015). About MySQL, available: https://www.mysql.com/about/ [accesses 2 November 2015].
46. Steve Suehring. (2002). MySQL Bible, Wiley Publishing, Inc., [online] available: http://www.chettinadtech.ac.in/g_article/Textbook2%20%20MySQL%20Bible.pdf [accessed 2 November 2015].
47. Daniel Bartholomew. (2012). MariaDB vs. MySQL. MariaBD, [online], September, available: http://www.eandbsoftware.org/wp-content/uploads/2015/03/MariaDB_vs_MySQL_-_MariaDB_White_Paper_-_08_26_13_001.pdf [accessed 2 November 2015].
48. Sentiment viz (2013). Tweet Sentiment Visualisation, available: https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/ [accessed 03 November 2015].
49. Socialmention (2008). Socialention application, available: http://www.socialmention.com/ [accessed 03 November 2015].