**Abstract**

Users receive information which must be filtered, processed, analysed, reviewed, consolidated and distributed or acted upon. Prior art tools automatically processing content to assign sentiment to the content are ineffective as essential aspects such as context are not considered. Embodiments of the invention provide automatic contextual based sentiment classification of content in terms of both sentiments expressed and their intensity. Further a content set is analysed to rapidly establish an “at-a-glance” type assessment of the key topics/themes present within the content set and sentimentally annotate each. Importantly embodiments of the invention also provide for a user to establish the basis for the sentiment associated with an item of or set of content, i.e., to make it explainable. Further embodiments of the invention provide for the establishment of psychological tone to sentiments where the sentiments and psychological tones to be tuned from the context or domain of the content.

**Diagram**

- **Term 410**
  - Negative Sentiment Seed Terms 420
  - Positive Sentiment Seed Terms 430
  - Target Domain Document Set 440
  - Negative Sentiment Seed Co-Occurrence Counts 450
  - Positive Sentiment Seed Co-Occurrence Counts 460

- **Determine Sentiment Label of Term 470**
- **Sentiment Lexicon 480**
- **Report Sentiment 475**
Figure 5

500 Select Document Set

510 Identify Core Multi-Document Concepts

515 Filter Identified Core Multi-Document Concepts

520 Select Core Multi-Document Concept

525 Method?

530 Document Summary

Determine Document Count Based Sentiment for Core Concept

535 Last

No

Yes

540 Generate Results for Document Count Based Sentiment of Selected Core Concept

545 Present to User Ranked List of Concepts to Show Concepts Ranked Most Important Along With Indicator of Term Sentiment

550 All Occurrences

Determine Context Count Based Sentiment by Identifying Each Instance of Concept in Each Document Within Set
METHOD AND SYSTEM RELATING TO SENTIMENT ANALYSIS OF ELECTRONIC CONTENT

CROSS-REFERENCE TO RELATED APPLICATIONS


FIELD OF THE INVENTION

[0002] The present invention relates to published content and more specifically to the processing of published content for users to associate sentiment to the content.

BACKGROUND OF THE INVENTION

[0003] In 2008, Americans consumed information for approximately 1.3 trillion hours, or an average of almost 12 hours per day per person (Global Information Industry Center, University of California at San Diego, January 2010). Consumption totaled 3.6 zettabytes (3.6x10^21 bytes) and 10,845 trillion (10,845x10^15) words, corresponding to 100,000 words and 34 gigabytes for an average person on an average day. This information comes from over twenty different sources of information, from newspapers and books through to online media, social media, satellite radio, and Internet video although the traditional media of radio and TV still dominated consumption per day.

[0004] Computers and the Internet have had major effects on some aspects of information consumption. In the past, information consumption was overwhelmingly passive, with telephone being the only interactive medium. However, with computers, a full third of words and more than half of digital data are now received interactively. Reading, which was in decline due to the growth of television, tripled from 1980 to 2008, because it is the overwhelmingly preferred way to receive words on the Internet. At the same time portable electronic devices and the Internet have resulted in a large portion of the population in the United States for example becoming active generators of information throughout their daily lives as well as active consumers augmenting their passive consumption. Social media such as Facebook™ and Twitter™, blogs, website comment sections, Bing™, Yahoo!™ have all contributed in different ways to the active generation of information by individuals which augments that generated by enterprises, news organizations, Government, and marketing organizations.

[0005] Globally the roughly 27 million computer servers active in 2008 processed 9.57 zettabytes of information (Global Information Industry Center, University of California at San Diego, April 2011). This study also estimated that enterprise server workloads are doubling about every two years and whilst a substantial portion of this information is incredibly transient overall the amount of information created, used, and retained is growing steadily.

[0006] The exploding growth in stored collections of numbers, images and other data represents one facet of information management for organizations, enterprises, Governments and individuals. However, even what was once considered “mere data” becomes more important when it is actively processed by servers as representing meaningful information delivered for an ever-increasing number of uses.

Overall the 27 million computer servers were estimated as providing an average of 3 terabytes of information per year to each of the estimated 3.18 billion workers in the world’s labor force.

[0007] Increasingly, a corporation’s competitiveness hinges on its ability to employ innovative search techniques that help users discover data and obtain useful results. In some instances automatically offering recommendations for subsequent searches or extracting related information are beneficial. To gain some insight into the magnitude of the problem consider the following:

[0008] In 2009 around 3.7 million new domains were registered each month and as of June 2011 this had increased to approximately 4.5 million per month;

[0009] approximately 45% of Internet users are under 25;

[0010] there are approximately 600 million wired and 1,200 million wireless broadband subscriptions globally;

[0011] approximately 85% of wireless handsets shipped globally in 2011 included a web browser;

[0012] there are approximately 2.1 billion Internet users globally with approximately 2.4 billion social networking accounts;

[0013] there are approximately 800 million users on Facebook™ and approximately 225 million Twitter™ accounts;

[0014] there are approximately 250 million tweets per day and approximately 250 million Facebook™ activities;

[0015] there are approximately 3 billion Google™ searches and 300 million Yahoo™ searches per day.

[0016] Accordingly it would be evident that users face an overwhelming barrage of information (content) that must be filtered, processed, analysed, reviewed, consolidated and distributed or acted upon. For example a market researcher seeking to determine the perception of a particular product may wish to rapidly collate sentiments from reviews sourced from websites, press articles, and social media.

[0017] Similarly, a search by a user using the terms “Barack Obama Afghanistan” with Google™ run on May 2, 2012 returns approximately 524 million “hits” in a fraction of a second. These are displayed, by default in the absence of other filters by the user, in an order determined by rules executed by Google™ servers relating to factors including, but not limited to, match to user entered keywords and the number of times a particular webpage or item of content has been opened. However, within this search the same content may be reproduced multiple times in different sources legitimately as well as having been plagiarized partially into other sources as well as the same event being presented through different content on other websites. Accordingly, different occurrences of Barack Obama visiting Afghanistan or different aspects of his visit to Afghanistan may become buried in an overwhelming reporting of his last visit or the repeated occurrence of strategic photo opportunities during the visit during a campaign.

[0018] Accordingly, it would be beneficial for the user to be able to retrieve a collection of multiple items of content, commonly referred to as documents, which mention one or more concepts or interests, and automatically cluster them into cohesive groups that relate to the same concepts or interests. Each cohesive group (or cluster) formed thereby consists of one or more documents from the original collection which describe the same concept or interest even where the docu-
ments have perhaps a different vocabulary. Even when a user identifies an item of content of interest, for example a review of a product, then the salient text may be buried within a large amount of other content or alternatively the item of content may be formatted for display upon laptops, tablet PCs, etc, whereas the user is accessing the content on a portable electronic device such as a smartphone or portable gaming console for example.

[0019] Accordingly it would be beneficial for the user to be able to access the salient text contained in one or more items of content, based on learned semantic and content structure cues so that extraneous elements of the item of content are removed. Accordingly it would be beneficial to provide a tool for inducing content scraping automatically to filter content to that necessary or automatically extracting core text for viewing on constrained screen devices or vocalizing through a screen reader. Automated summarization or text simplification may also form extensions of the scraper.

[0020] Other aspects and features of the present invention will become apparent to those ordinarily skilled in the art upon review of the following description of specific embodiments of the invention in conjunction with the accompanying figures.

SUMMARY OF THE INVENTION

[0021] It is an object of the present invention to provide improvements in the art relating to published content and more specifically to the processing of published content for users to associate sentiment to content, cluster content for review, and extract core text.

[0022] In accordance with an embodiment of the invention there is provided a method comprising:

[0023] receiving an item of content;

[0024] parsing the item of content with a microprocessor to generate a linguistic annotated item of content with language associations;

[0025] retrieving from a term selection rules repository stored upon a memory at least a rule of a plurality of rules;

[0026] applying with the microprocessor the at least a rule of the plurality of rules to establish a set of candidate sentiment carrying terms within the linguistic annotated item of content;

[0027] querying the set of candidate sentiment carrying terms against a target-domain sentiment lexicon to generate a set of sentiment labeled terms; and

[0028] applying to the linguistic annotated item of content a set of sentiment labeling rules established in dependence at least the set of sentiment labeled terms to generate a sentiment label for the item of content.

[0029] In accordance with an embodiment of the invention there is provided a method comprising:

[0030] a) receiving an item of content;

[0031] b) receiving upon a microprocessor an indication of a predetermined portion of the item of content to analyze;

[0032] c) establishing with the microprocessor a plurality of positive sentiment terms and a plurality of negative sentiment terms;

[0033] d) parsing with the microprocessor the predetermined portion of the item of content to count occurrences of a positive sentiment term of the plurality of positive sentiment terms to establish a positive sentiment count;

[0034] e) parsing with the microprocessor the predetermined portion of the item of content to count occurrences of a negative sentiment term of the plurality of negative sentiment terms to establish a negative sentiment count; and

[0035] f) determining with the microprocessor a sentiment label to associate with the item of content in dependence upon at least one of the occurrences of the positive sentiment term and occurrences of the negative sentiment term.

[0036] In accordance with an embodiment of the invention there is provided a method comprising:

[0037] processing with a microprocessor the item of content to determine occurrences of content sentiment-carrying terms;

[0038] displaying to a user the sentiment labels of content sentiment-carrying terms within the item of content; and

[0039] presenting to the user any sentiment intensity variation based on matching at least one of a predetermined sentence and a phrasal syntactic structure of the document with a repository of syntactic structure patterns.

[0040] In accordance with an embodiment of the invention there is provided a method comprising:

[0041] a) receiving a plurality of items of content;

[0042] b) identifying with a microprocessor within the plurality of items of content at least a core multi-item concept of a plurality of core multi-item concepts, each core multi-item concept relating to a concept contained at least within a predetermined portion of the plurality of items of concept;

[0043] c) selecting a core multi-item concept from the plurality of core multi-item concepts; and

[0044] d) establishing with the microprocessor a sentiment relating to the core multi-item concept for the plurality of items of content.

[0045] Other aspects and features of the present invention will become apparent to those ordinarily skilled in the art upon review of the following description of specific embodiments of the invention in conjunction with the accompanying figures.

BRIEF DESCRIPTION OF THE DRAWINGS

[0046] Embodiments of the present invention will now be described, by way of example only, with reference to the attached Figures, wherein:

[0047] FIG. 1A depicts a network accessible by a user and content sources accessible to the user with respect to embodiments of the invention;

[0048] FIG. 1B depicts an electronic device supporting communications and interactions for a user according to embodiments of the invention;

[0049] FIGS. 2A and 2B depict a machine based sentiment learning and classification process according to the prior art;

[0050] FIG. 3 depicts a flowchart of a process for a sentiment classification process using a target-domain sentiment lexicon according to an embodiment of the invention;

[0051] FIG. 4 depicts a flowchart of a process for a target domain sentiment lexicon generation process according to an embodiment of the invention; and

[0052] FIG. 5 depicts a process flow for associating key concepts within multiple documents and associating sentiments to the key concepts according to an embodiment of the invention.
DETAILED DESCRIPTION

[0053] The present invention is directed to published content and more specifically to the processing of published content for users to associate sentiment to content, cluster content for review, and extract core text.

[0054] The ensuing description provides exemplary embodiment(s) only, and is not intended to limit the scope, applicability or configuration of the disclosure. Rather, the ensuing description of the exemplary embodiment(s) will provide those skilled in the art with an enabling description for implementing an exemplary embodiment. It being understood that various changes may be made in the function and arrangement of elements without departing from the spirit and scope as set forth in the appended claims.

[0055] A “portable electronic device” (PED) as used herein and throughout this disclosure, refers to a wireless device used for electronic communications that requires a battery or other independent form of energy for power. This includes devices, but is not limited to, such as a cellular telephone, smartphone, personal digital assistant (PDA), portable computer, pager, portable multimedia player, portable gaming console, laptop computer, tablet computer, and an electronic reader. A “fixed electronic device” (FED) as used herein and throughout this disclosure, refers to a wired or wireless device used for electronic communications that may be dependent upon a fixed source of power, employ a battery or other independent form of energy for power. This includes devices, but is not limited to, such as a portable computer, personal computer, Internet enabled display, gaming console, computer server, kiosk, and a terminal.

[0056] A “network operator/service provider” as used herein may refer to, but is not limited to, a telephone or other company that provides services for mobile phone subscribers including voice, text, and Internet; telephone or other company that provides services for subscribers including but not limited to voice, text, Voice-over-IP, and Internet; a telephone, cable or other company that provides wireless access to local area, metropolitan area, and long-haul networks for data, text, Internet, and other traffic or communication sessions; etc.

[0057] “Content”, “input content” and/or “document” as used herein and through this disclosure refers to an item or items of information stored electronically and accessible to a user for retrieval or viewing. This includes, but is not limited to, documents, images, spreadsheets, databases, audiovisual data, multimedia data, encrypted data, SMS messages, social media data, data formatted according to a markup language, and information formatted according to a portable document format.

[0058] A “web browser” as used herein and through this disclosure refers to a software application for retrieving, presenting, and traversing information resources on the World Wide Web identified by a Uniform Resource Identifier (URI) and may be a web page, image, video, or other piece of content. The web browser also allows a user to access and implement hyperlinks present in accessed resources to navigate their browsers to related resources. A web browser may also be defined within the scope of this specification as an application software or program designed to enable users to access, retrieve and view documents and other resources on the Internet as well as access information provided by web servers in private networks or files in file systems.

[0059] An “application” as used herein and through this disclosure refers to a software application, also known as an “app”, which is computer software designed to help the user to perform specific tasks. This includes, but is not limited to, web browser, enterprise software, accounting software, information work software, content access software, education software, media development software, office suites, presentation software, work processing software, spreadsheets, graphics software, email and blog client software, personal information systems and desktop publishing software. Many application programs deal principally with multimedia, documentation, and/or audiovisual content in conjunction with a markup language for annotating a document in a way that is syntactically distinguishable from the content. Applications may be bundled with the computer and its system software, or may be published separately.

[0060] A “user,” as used herein and through this disclosure refers to, but is not limited to, a person or device that generates, receives, analyses, or otherwise accesses content stored electronically within a portable electronic device, fixed electronic device, network accessible server, or other source storing content.

[0061] A “server” as used herein and through this disclosure refers to a computer program running to serve the requests of other programs, the “clients”. Thus, the “server” performs some computational task on behalf of “clients” which may either run on the same computer or connect through a network. Accordingly such “clients” therefore being applications in execution by one or more users on their PED/FED or remotely at a server. Such a server may be one or more physical computers dedicated to running one or more services as a host. Examples of a server include, but are not limited to, database server, file server, mail server, print server, and web server.

[0062] Referring to FIG. 1A there is depicted a network supporting communications and interactions between devices connected to the network and executing functionalities according to embodiments of the invention with a first and second user groups 100A and 1000B respectively to a telecommunications network 100. Within the representative telecommunication architecture a remote central exchange 180 communicates with the remainder of a telecommunication service providers network via the network 100 which may include for example long-haul OC-48/OC-192 backbone elements, an OC-48 wide area network (WAN), a Passive Optical Network, and a Wireless Link. The remote central exchange 180 is connected via the network 100 to local, regional, and international exchanges (not shown for clarity) and therein through network 100 to first and second wireless access points (AP) 120 and 110 respectively which provide Wi-Fi cells for first and second user groups 100A and 100B respectively.

[0063] Within the cell associated with first AP 120 the first group of users 100A may employ a variety of portable electronic devices (PEDs) including for example, laptop computer 155, portable gaming console 135, tablet computer 140, smartphone 150, cellular telephone 145 as well as portable multimedia player 130. Within the cell associated with second AP 110 the second group of users 100B may employ a variety of portable electronic devices (not shown for clarity) but may also employ a variety of fixed electronic devices (PEDs) including for example gaming console 125, personal computer 115 and wireless/Internet enabled television 120 as well as cable modem 105 which links second AP 110 to the network 100.

[0064] Also connected to the network 100 is cell tower 125 that provides, for example, cellular GSM (Global System for
Mobile Communications) telephony services as well as 3G and 4G evolved services with enhanced data transport support. Cell tower 125 provides coverage in the exemplary embodiment to first and second user groups 100A and 100B. Alternatively the first and second user groups 100A and 100B may be geographically disparate and access the network 100 through multiple cell towers, not shown for clarity, distributed geographically by the network operator or operators. Accordingly, the first and second user groups 100A and 100B may access their particular communications interfaces to communicate to the network 100 through one or more communications standards as, for example, IEEE 802.11, IEEE 802.15, IEEE 802.16, IEEE 802.20, UMTS, GSM 850, GSM 900, GSM 1800, GSM 1900, GPRS, ITU-R 5.138, ITU-R 5.150, ITU-R 5.280, and IMT-2000. It would be evident to one skilled in the art that many portable and fixed electronic devices may support multiple wireless protocols simultaneously, such that for example a user may employ various network access such as telephone and SMS and Wi-Fi/ WiMAX data transmission, VOW and Internet access.

[0065] Also communicated to the network 100 are first and second servers 110A and 110B respectively which host according to embodiments of the invention multiple services associated with content from one or more sources including, for example, but not limited to:

- social media 160 such as Facebook™, Twitter™, and LinkedIn™, etc;
- web feeds 165 such as formatted according to RSS and/or Atom formats to publish frequently updated works;
- web portals 170 such as Yahoo™, Google™, Baidu™, and Microsoft’s Bing™ for example;
- broadcasters 175 including Fox, NBC, CBS, and Comcast for example who provide content via multiple media including for example satellite, cable, and Internet;
- print media 180 including for example USA Today, Washington Post, Los Angeles Times and China Daily;
- websites 185 including, but not limited to, manufacturers, market research, consumer research, newspapers, journals, and financial institutions.

[0072] Also connected to network 100 is application server 105 which provides software system(s) and software application(s) associated with receiving retrieved content and processing said published content for users to associate sentiment to content, cluster content for review, and extract core text as discussed below in respect of embodiments of the invention. First and second servers 110A and 110B and application server 105 together with other servers not shown for clarity may also provide dictionaries, speech recognition software, product databases, inventory management databases, retail pricing databases, shipping databases, customer databases, software applications for downloading to fixed and portable electronic devices, as well as Internet services such as a search engine, financial services, third party applications, directories, mail, mapping, social media, news, user groups, and other Internet based services.

[0073] Referring to FIG. 1B there is depicted an electronic device 1004, supporting communications and interconnections according to embodiments of the invention with local and/or remote services. Electronic device 1004 may be for example a PED, PEB, a terminal, or a kiosk. Also depicted within the electronic device 1004 is the protocol architecture as part of a simplified functional diagram of a system 1000 that includes an electronic device 1004, such as a smartphone 155, an access point (AP) 1006, such as first Wi-Fi AP 110, and one or more remote servers 1007, such as communication servers, streaming media servers, and routers for example such as first and second servers 110A and 110B respectively. Remote server cluster 1007 may be coupled to AP 1006 via any combination of networks, wired, wireless and/or optical communication links such as discussed above in respect of FIG. 1. The electronic device 1004 includes one or more processors 1010 and a memory 1012 coupled to processor(s) 1010. AP 1006 also includes one or more processors 1011 and a memory 1013 coupled to processor(s) 1011. A non-exhaustive list of examples for any of processors 1010 and 1011 includes a central processing unit (CPU), a digital signal processor (DSP), a reduced instruction set computer (RISC), a complex instruction set computer (CISC) and the like. Furthermore, any of processors 1010 and 1011 may be part of application specific integrated circuits (ASICs) or may be a part of application specific standard products (ASSPs). A non-exhaustive list of examples for memories 1012 and 1013 includes any combination of the following semiconductor devices such as registers, latches, ROM, EEPROM, flash memory devices, non-volatile random access memory devices (NVRAM), SDRAM, DRAM, double data rate (DDR) memory devices, SRAM, universal serial bus (USB) removable memory, and the like.

[0074] Electronic device 1004 may include an audio input element 1014, for example a microphone, and an audio output element 1016, for example a speaker, coupled to any of processors 1010. Electronic device 1004 may include a video input element 1018, for example a video camera, and a video output element 1020, for example an LCD display, coupled to any of processors 1010. Electronic device 1004 includes one or more applications 1022 that are typically stored in memory 1012 and are executable by any combination of processors 1010. Electronic device 1004 includes a protocol stack 1024 and AP 1006 includes a communication stack 1025. Within system 1000 protocol stack 1024 is shown as IEEE 802.11 protocol stack but alternatively may exploit other protocol stacks such as an Internet Engineering Task Force (IETF) multimedia protocol stack for example. Likewise AP stack 1025 exploits a protocol stack but is not expanded for clarity. Elements of protocol stack 1024 and AP stack 1025 may be implemented in any combination of software, firmware and/or hardware. Protocol stack 1024 includes an IEEE 802.11-compatible PHY module 1026 that is coupled to one or more Front-End Tx/Rx & Antenna 1028, an IEEE 802.11-compatible MAC module 1030 coupled to an IEEE 802.2-compatible LLC module 1032. Protocol stack 1024 includes a network layer IP module 1034, a transport layer User Datagram Protocol (UDP) module 1036 and a transport layer Transmission Control Protocol (TCP) module 1038.

[0075] Protocol stack 1024 also includes a session layer Real Time Transport Protocol (RTP) module 1040, a Session Announcement Protocol (SAP) module 1042, a Session Initiation Protocol (SIP) module 1044 and a Real Time Streaming Protocol (RTSP) module 1046. Protocol stack 1024 includes a presentation layer media negotiation module 1048, a call control module 1050, one or more audio codecs 1052 and one or more video codecs 1054. Applications 1022 may be able to create maintain and/or terminate communication sessions with any of remote servers 1007 by way of AP 1006. Typically, applications 1022 may activate any of the SAP, SIP,
RTSP, media negotiation and call control modules for that purpose. Typically, information may propagate from the SAP, SIP, RTSP, media negotiation and call control modules to PHY module 1026 through TCP module 1038, IP module 1034, LLC module 1032 and MAC module 1030.

[0076] It would be apparent to one skilled in the art that elements of the PED 1004 may also be implemented within the AP 1006 including but not limited to one or more elements of the protocol stack 1024, including for example an IEEE 802.11-compatible PHY module, an IEEE 802.11-compatible MAC module, and an IEEE 802.2-compatible LLC module 1032. The AP 1006 may additionally include a network layer IP module, a transport layer User Datagram Protocol (UDP) module and a transport layer Transmission Control Protocol (TCP) module as well as a session layer Real Time Transport Protocol (RTP) module, a Session Announcement Protocol (SAP) module, a Session Initiation Protocol (SIP) module and a Real Time Streaming Protocol (RTSP) module, media negotiation module, and a call control module.

[0077] As depicted remote server cluster 1007 comprises a firewall 1007A through which the discrete servers within the remote server cluster 1007 are accessed. Alternatively remote server 1007 may be implemented as multiple discrete independent servers each supporting a predetermined portion of the functionality of remote server cluster 1007. As presented the discrete servers include application servers 10073 dedicated to running certain software applications, communications server 1007C providing a platform for communications networks, database server 1007D providing database services to other computer programs or computers, web server 1007E providing HTTP clients connectivity in order to send commands and receive responses along with content, and proxy server 1007F that acts as an intermediary for requests from clients seeking resources from other servers.

[0078] Contextual Sentiment Classification:

[0079] Prior Art:

[0080] Within the prior art multiple approaches to classifying or assigning a sentiment for an item of content, typically a document or portion of a document, exist. However, these existing sentiment filtering approaches simply determine occurrences of a keyword with positive and negative terms to establish an overall sentiment. However, this analysis does not provide any context in respect of these occurrences with their context. As outlined above the phrase “last night I drove to see Terminator 3 in my new Fiat 500, after eating at Stonewall, the truffle bison burger was great” would be interpreted as positive feedback even though the positive term is associated with the food rather than either the film “Terminator 3” or the vehicle “Fiat 500.” Accordingly, it would be beneficial for sentiment analysis of content to be contextually aware.

[0081] Referring to FIGS. 2A and 2B there are depicted first and second schematic representations 200 and 2000 respectively of the prior art of Pang et al for sentiment classification, which employs the classic ‘bag-of-words’ feature representation for machine learning classification. Referring to first schematic 200 there is depicted a first stage of the prior art process wherein a learning process is performed. A training dataset 205 is stored upon a server for example wherein the training dataset 205 comprises a predetermined set of documents that serve as training examples for the prior art process wherein typically half of the training document set 205 are labelled as expressing positive sentiment, and the other half of the training document set 205 are labelled as expressing negative sentiment. The training document set 205 are then parsed in a feature vocabulary extraction process 210 to provide a unique set of words found in the training document set 205. Optionally these are stored with associated frequency counts. The “feature vocabulary list” extracted in feature vocabulary extraction process 210 is then optionally reduced through feature engineering 220 to a smaller set via thresholds which may for example be based on word frequencies, chi-squared distribution (also known as chi-square or $\chi^2$ distribution), or information theoretic means for example. New features may also be introduced via documents or corpus analysis. The training document set 205 are then processed using a standard machine learning algorithm 230, such as for example Naïve Bayes, Support Vector Machines, and Maximum Entropy to generate a classification model 235 based on the association of provided features to the document sentiment labels.

[0082] Now referring to second schematic 2000 a second stage of the prior art is depicted wherein an input document 240 is to be analyzed for sentiment. A feature vocabulary 245 was used to generate a sentiment classification model 255 as discussed above in respect of first schematic 200 during a machine learning training process 230. Accordingly the input document 240 is processed by an initial document feature engineering 250 process which converts the input document 240 to a format that matches the features employed in the sentiment classification process 260 which is based upon a machine learning model 255. This transformation follows the same process as feature engineering 220 in first schematic 200 of FIG. 2A. Accordingly the sentiment classification process 260 assigns a sentiment label to the features derived from the input document 240 wherein the positive or negative sentiment is output as document sentiment label 270 and associated with the input document 240.

[0083] Such prior art approaches suffer from a number of serious limitations, which are addressed by embodiments of the current invention. The limitations include the fact that the sentiment label 270 applied to an input document 240 is not explainable. Most machine-learning based classification systems generate an opaque high-dimensional model such that the sentiment label associated with a document cannot be mapped back to the document, and thus there is no easily understandable method to describe how the class-association statistics associated with individual features are used to derive the sentiment label. This “black-box” nature of the machine learning classifier can unnerve those who depend professionally on the veracity of the sentiment label to make business decisions.

[0084] Additionally the performance of these supervised machine learning techniques is dependent on the degree to which the training data set and testing data match with respect to domain, topic and time-period. However, it would be evident that a term may provide positive or negative sentiment and accordingly should not form part of the feature vocabulary. For example the word “conservative” may be considered to have positive sentiment in content from the financial domain, but may have negative sentiment in content relating to movie reviews or an artistic genre. Accordingly prior art machine learning based solutions do not ensure that the sentiment associated with a document’s constituent terms is derived from the same sentiment context as the document. Without this domain match, highly descriptive words in testing or production document may have a different sentiment
than those given in the training document set. Prior art techniques are also not arrived at by a rigorous linguistic analysis of the document.

[0085] It would also be evident that the prior art machine learning classification approaches can only operate on information that they have encountered before, i.e. only those features are supported that were included in the training document set’s vocabulary. Occurrences of “unseen” words, i.e. words not within the training document set which are extracted into the feature vocabulary set, are essentially ignored. Another limitation within prior art techniques is the ability to classify small documents, especially data sets derived from cellular SMS messages or Twitter status updates for example, as these documents are too small to accurately be classified by machine learning based sentiment classification. However, in many instances such documents are desirable as the focus of sentiment classification as a substantial negative or positive sentiment across SMS messages, Tweets, or Facebook status updates provide rapid near real-time analysis of an event or occurrence. For example, a broadcaster upon broadcasting a potentially controversial episode or program may gauge their viewers’ responses as the broadcast progresses and track the subsequent evolution of demographic breakdowns in sentiment or evolution of consensus for example.

[0086] Contextual Sentiment Classification—Sentiment Classification Process:

[0087] The contextual sentiment classification of content according to embodiments of the invention is achieved through use of two core processes. These are a sentiment classification process which exploits a target-domain sentiment lexicon and generation of the target-domain sentiment lexicon. Referring to FIG. 3 there is presented an overview process flowchart 300 according to an embodiment of the invention by which an input document 310 is labelled with a sentiment label 370 as an output of the overview process flowchart 300 class, with optional sentiment intensity, via a linguistic parser 320, term selection rules 340, target-domain sentiment lexicon 350, and document sentiment labelling rules 380. The sentiment label 370 being generated in dependence of one or more sentiment labelled terms 360 generated through the process.

[0088] Accordingly the process begins with input content, document 310, which is transformed via a parser 320 into an annotated form with associations including, but not limited to, part-of-speech, phrasal chunks, and grammatical relations associated with terms that constitute the input document 310. Rules retrieved from a term selection rules repository 340 are then employed to derive a set of candidate sentiment carrying terms, selected terms 330, from the annotated version of the document 310 generated by parser 320. Each selected term 330 is then queried in a target-domain sentiment lexicon 350 to create a list of terms, the sentiment labelled terms 360, with associated sentiment labels and optionally associated sentiment intensity. These sentiment labelled terms 360 with any associated elements are then employed with the linguistic annotated version of the document generated by the parser 320 to apply a set of document sentiment labeling rules 380 in order to generate a document sentiment label 370. Similarly optionally associated sentiment intensities can be employed in conjunction with the document sentiment labelling rules 380 to establish an optional sentiment intensity level for the document 310.

[0089] Optionally, the sentiment labelled terms 360, have associated with them one or more sentiment labels and optionally one or more associated sentiment intensities. For example, the term “git” may have the sentiment label of “hate” associated with an intensity of “weak” whereas “loathe” may have the same sentiment label of “hate” but an intensity of “extreme.” It would be evident to one skilled in the art that the target-domain sentiment lexicon 350 may established in dependence upon the domain of the input content, document 310. The domain may be one or more fields, the fields including but not limited to, an area of human activity, an area of human interest, an area of human endeavor, a topic, a subject, an area of academic interest, an area of academic specialization, a profession, an aspect of business, an aspect of entertainment, and an aspect of personal relationships. The term selection rules repository 340 and the rules stored within it may optionally be established upon the domain of the input content or alternatively these may be established in dependence upon one or more factors including the enterprise/service provider executing the sentiment classification process, the software system and/or software system provider supplied repository and rules, user preferences, and preferences of a requestor of a sentiment analysis.

[0090] It would be evident to one skilled in the art that the process described above in respect of FIG. 3 may be applied to a plurality of documents to form the input content wherein the results of each of the plurality of documents may be reported individually or the results may be collated to provide a single determined sentiment or an analysis such as numbers expressing strong positive, positive, mildly positive, neutral, mildly negative, negative, and strong negative sentiment. Such analysis may include optionally reporting events of particular sentiments with intense or very strong sentiment. Optionally, the results of a sentiment analysis such as described supra may be employed in other processes, such as, for example, where the sentiment labelled terms become elements of core text to be extracted from a document through a salient content extraction process such that the result of such a process is a document or documents being reduced to the text associated with the sentiment labelled terms.

[0091] Contextual Sentiment Classification—Target-Domain Sentiment Lexicon Generation Process:

[0092] As noted supra the sentiment classification process exploits a target-domain sentiment lexicon and accordingly the generation of the target-domain sentiment lexicon, which is a separate process is described here. Referring to FIG. 4 there is illustrated a process flowchart schematic 400 wherein an input term 410 is assigned a target-domain sentiment label with a sentiment lexicon 480, with an optional sentiment intensity, by analyzing the co-occurrence counts of this input term 410 with negative sentiment seed terms 420 and positive sentiment seed terms 430 in a target-domain document set 440.

[0093] The process flowchart schematic 400 depicting the lexicon generation process is based upon a determination process. This process is based upon generating two counts, the first count being of documents in the target-domain document set 440 containing both an input term 410 and one or more negative sentiment seed terms of the set of negative sentiment seed terms 420 and storing this negative sentiment seed co-occurrence count 450. The second count being of documents in the target-domain document set 440 containing both an input term 410 and one or more positive sentiment seed terms of the set of positive sentiment seed terms 430 and
is stored as the positive sentiment seed co-occurrence count 460. Optionally, the co-occurrence counts, being negative sentiment seed co-occurrence count 450 and positive sentiment seed co-occurrence count 460, may count co-occurrences in one or more of paragraphs, sentences, sliding windows of word (optionally truncated by sentence end punctuations), and via grammatical relations.

[0094] The counts of negative and positive seed term co-occurrence counts 450 and 460 respectively are analyzed to determine the target-domain sentiment label of the term, the sentiment label of term 470. Subsequently the input term, sentiment label, and (optionally) count information, is reported to a user as shown in the process by Report Sentiment 475 and is also stored into a target-domain sentiment lexicon 480. The analysis and determination of the sentiment label of term 470 may for example simply be the higher score if the negative term counts, negative sentiment seed co-occurrence count 450, are approximately equal the positive term counts, positive sentiment seed co-occurrence count 460. Alternatively, if the classes are imbalanced the analysis may involve a normalization step to reduce the weighting of the more frequent class or terms within each of the negative and positive seed term co-occurrence counts 450 and 460 respectively may have weightings associated with them such that certain terms if occurring in a document have higher weight than others.

[0095] It would be evident that input term 410 may be an item of content without any prior consideration or analysis and hence may be an item of content retrieved from one or more sources as discussed above in respect of FIG. 1 or may be an item of content received in real time such that for example Twitter tweets or Facebook posts may be analysed as they are published thereby allowing an organization the ability to monitor sentiments in essentially real-time. It would also be evident that the item of content may be a single document, such as for example a marketing report or a customer comment received online; a collection of documents; a webpage such as for example a blog, a reporters column, a competitor’s product, a consumer organization’s report; or a web domain such that all content within the web domain is analysed such as for example web domains for consumer organizations, newspapers, magazines, competitors, and retailers. It would be further evident that input term 410 may be initially filtered for an occurrence of a particular keyword, subset of a set of keywords, or all keywords in a set of keywords. Optionally the content may also be processed such that locations of the negative and positive sentiment seed terms relative to one or more keywords are determined and only those meeting a predetermined threshold condition are counted into the respective negative and positive sentiment seed co-occurrence counts.

[0096] The content in addition to a social network status update may therefore as discussed and presented supra include, but not be limited to, other content such as an email, a news article, a blog post, a forum comment, a stock report, a news cast, a web page, or any other form of user generated content and/or content generated from an editorial process. The document may have a structure, such as for example including a title, body, and summary, with one or more paragraphs. The structure could be in the form of a template or a frame. Accordingly sentiment analysis may be performed on these structural elements independently to provide multiple sentiments for the item of content or be combined with a weighting in dependence of the structure to provide a sentiment for the content overall. For example, sentiments within the title and summary may be weighted higher than those within the body of the content.

[0097] Optionally, according to another embodiment of the invention a domain-detection component may be provided which identifies the domain of an input document, and employs this domain-identification tag to choose one (or more) target-domain sentiment lexicons from a plurality of stored lexicons. According to another embodiment of the invention a sentiment may be provided with an ordinal scale, for example from \{0,1\}, \{-1,1\}, \{-2,4.5\}, or \{-5,5\}.

[0098] In another embodiment of the invention in addition to the sentiment label for the document, a set of sentiment labels, with optional intensity metrics, could be provided for each constituent term in the document. Optionally the sentiment returned for the document could also contain psychological tone qualifications, such as anger, affinity, disgust, sorrow, etc. based upon exploiting known emotion and attitude ontologies.

[0099] The invention could also be combined with a display method which can show the document and the associated sentiment, with optional annotations on selected lexical units that serve to explain the sentiment provided thereby.

[0100] Accordingly, advantages of embodiments of the invention include:

[0101] providing improved sentiment analysis as the sentiment generated is based on a targeted-domain sentiment lexicon.

[0102] domain-independent sentiment analysis can be provided when a contextual sentiment analysis system is coupled with a large sample of documents that pertain to a plurality of subjects of interest to a variety of readers; and

[0103] ability to describe why a sentiment label has been applied to a document by providing the underlying sentiment(s) associated with selected terms in the document;

[0104] a parser is employed to select the salient terms from the document thereby allowing the system to assign sentiment to only the relevant sentiment-carrying terms.

[0105] It would be evident that beneficially the parser allows for identification of the syntactic and semantic linguistic roles of the terms that constitute the document being analyzed for sentiment. Further by employing a set of document sentiment labeling rules, that operate on the syntactic, semantic and sentiment meta-data associated with the terms constituting a document, embodiments of the invention can generate a sentiment based on the linguistic structure of the document, rather than employing the prior art linguistic-structure-berief ‘bag-of-words’ machine learning sentiment analysis framework.

[0106] Contextual Sentiment Classification—Multi-Document Key Concept Generation and Sentiment Association Process:

[0107] Referring to FIG. 5 there is depicted a process flowchart 500 according to an embodiment of the invention for associating key concepts within multiple documents and associating sentiments to the key concepts. As depicted process flowchart 500 begins at step 505 wherein the document set is selected by one or more methods including, but not limited to, manual selection by the user, automatically by an application in execution associated with the user, automatically by an application in execution upon a software system associated with a service subscribed to by the user, and an
application in execution upon a software system associated with a software application employed by the user. The process then proceeds to step 510 wherein the core multi-document concepts are identified. These core multi-document concepts being identified, for example, using a ranking technique including, but not limited to, frequency-based ranking, chi-square, mutual information, k-means clustering, vector-space centroids. The process then proceeds to step 515 wherein the list of key concepts may be filtered to reduce the derived, optionally ranked list, via one or more techniques including, but not limited to, threshold based cutoff, top predetermined number, confidence scores or by comparing with a stop-word list which consists of terms to be excluded as key concepts.

In step 520 the core multi-document concept is selected, e.g. highest ranking, wherein the process proceeds to step 525 for a determination as to the method to be employed is made, which are shown as “Document Summary” and “All Occurrences”. If “Document Summary” is selected, for example by the user, via a preference within the software application and/or software system, number of documents, and in dependence upon the core multi-document concept, then the process proceeds to step 530 wherein a document based sentiment for the given key concept is obtained for a document within the document set. In step 535 the process determines whether all documents within the document set have had document based sentiments established wherein the process loops back to step 530 if further documents remain or proceeds to step 540 wherein counts are generated for the positive, negative and neutral sentiments establishing how many documents for that sentiment it is the overall. Then in step 545 the user is presented with the category with the largest sentiment count, or alternatively is presented with the results for all three categories. The largest sentiment count category may then be employed according to embodiments of the invention for a variety of subsequent processes, such as for example rewarding customers within that category for their feedback which may be in some instances negative feedback but avoiding automatic rewarding for good feedback may result in a more honest feedback. Alternatively, the sentiment result may be employed to trigger other activities or events such as searching for that sentiment within a new document set.

If in step 525 the “All Occurrences” method was selected then the process proceeds to step 550 wherein the context-count-based sentiment for a given key concept is established by identifying the sentiment associated with each and every instance of the key concept as it occurs in each document being processed. Accordingly, the process then proceeds to step 545 again to present for example and an indicator that indicates the sentiment of the term based on the sentiment label derived using the results from step 550 via simple addition or through other sentiment classification techniques. The indication may for example be a colour coding, audiovisual coding, or another indicator as known within the art.

It would be evident that other statistical techniques and approaches may be employed in establishing the core multi-document concepts including identification by the user, identification by the software applications and/or software system using previously stored index terms, and entry of a search term and/or terms into a software application such as an Internet browser for example. Optionally, the filtering step 515 may be omitted or replaced with a user selection using a graphical user interface according to one or more techniques known in the prior art. As presented steps 525 through 550 of process flowchart 500 are depicted as occurring once for the top ranked core multi-document concept. However, it would be evident to one skilled in the art that these steps may be repeated for one or more of the core multi-document concepts resulting from the filtering step 515. For example, the top 5 concepts may be automatically processed or all concepts exceeding a threshold may be processed.

It would be evident that more or less categories may be established for the multi-document sentiment analysis of the sentiment set or that the process may be re-run once a particular overall sentiment has been assessed to refine the analysis, for example negative may be subsequently assessed for anger, frustration, calm for example. Within the embodiments of the invention a document within a document set may refer, for example, to an article, a blog, a social media post, an email, a comment posted to a website, a word processing document, an office document, a response to a survey, an item of multimedia content, and an item of audiovisual content. Optionally, the results from the process flowchart 500 relating to a sentiment analysis of a core concept or core concepts within a document set may be communicated through the software application or another software application, e.g. an electronic mail application, for distribution. According, a user may establish a sentiment analysis upon a software system and/or software application which periodically selects a predetermined number of documents to form a document set from a larger volume of documents and transmits the result of sentiment analysis and core concepts to the user such that for example a news service may not only identify the currently trending topics within any, Twitter™ but also automatically obtain associated with these the sentiment analysis.

Specific details are given in the above description to provide a thorough understanding of the embodiments. However, it is understood that the embodiments may be practiced without these specific details. For example, circuits may be shown in block diagrams in order not to obscure the embodiments in unnecessary detail. In other instances, well-known circuits, processes, algorithms, structures, and techniques may be shown without unnecessary detail in order to avoid obscuring the embodiments.

Implementation of the techniques, blocks, steps and means described above may be done in various ways. For example, these techniques, blocks, steps and means may be implemented in hardware, software, or a combination thereof. For a hardware implementation, the processing units may be implemented within one or more application specific integrated circuits (ASICs), digital signal processors (DSPs), digital signal processing devices (DSPDs), programmable logic devices (PLDs), field programmable gate arrays (FPGAs), processors, controllers, micro-controllers, microprocessors, other electronic units designed to perform the functions described above and/or a combination thereof.

Also, it is noted that the embodiments may be described as a process which is depicted as a flowchart, a flow diagram, a data flow diagram, a structure diagram, or a block diagram. Although a flowchart may describe the operations as a sequential process, many of the operations can be performed in parallel or concurrently. In addition, the order of the operations may be rearranged. A process is terminated when its operations are completed but could have additional steps not included in the figure. A process may correspond to a method, a function, a procedure, a subroutine, a subprogram,
Furthermore, embodiments may be implemented by hardware, software, scripting languages, firmware, middleware, microcode, hardware description languages and/or any combination thereof. When implemented in software, firmware, middleware, scripting language and/or microcode, the program code or code segments to perform the necessary tasks may be stored in a machine readable medium, such as a storage medium. A code segment or machine-executable instruction may represent a procedure, a function, a subprogram, a program, a routine, a subroutine, a module, a software package, a script, a class, or any combination of instructions, data structures and/or program statements. A code segment may be coupled to another code segment or a hardware circuit by passing and/or receiving information, data, arguments, parameters and/or memory contents. Information, arguments, parameters, data, etc. may be passed, forwarded, or transmitted via any suitable means including memory sharing, message passing, token passing, network transmission, etc.

For a firmware and/or software implementation, the methodologies may be implemented with modules (e.g., procedures, functions, and so on) that perform the functions described herein. Any machine-readable medium tangibly embodying instructions may be used in implementing the methodologies described herein. For example, software codes may be stored in a memory. Memory may be implemented within the processor or external to the processor and may vary in implementation where the memory is employed in storing software codes for subsequent execution to that when the memory is employed in executing the software codes. As used herein the term "memory" refers to any type of long term, short term, volatile, nonvolatile, or other storage medium and is not to be limited to any particular type of memory or number of memories, or type of media upon which memory is stored.

Moreover, as disclosed herein, the term "storage medium" may represent one or more devices for storing data, including read only memory (ROM), random access memory (RAM), magnetic RAM, core memory, magnetic disk storage mediums, optical storage mediums, flash memory devices and/or other machine readable mediums for storing information. The term "machine-readable medium" includes, but is not limited to portable or fixed storage devices, optical storage devices, wireless channels and/or various other mediums capable of storing, containing or carrying instruction(s) and/or data.

The methodologies described herein are, in one or more embodiments, performable by a machine which includes one or more processors that accept code segments containing instructions. For any of the methods described herein, when the instructions are executed by the machine, the machine performs the method. Any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine are included. Thus, a typical machine may be exemplified by a typical processing system that includes one or more processors. Each processor may include one or more of a CPU, a graphics-processing unit, and a programmable DSP unit. The processing system further may include a memory subsystem including main RAM and/or a static RAM, and/or ROM. A bus subsystem may be included for communicating between the components. If the processing system requires a display, such a display may be included, e.g., a liquid crystal display (LCD). If manual data entry is required, the processing system also includes an input device such as one or more of an alphanumeric input unit such as a keyboard, a pointing control device such as a mouse, and so forth.

The memory includes machine-readable code segments (e.g., software or software code) including instructions for performing, when executed by the processing system, one of more of the methods described herein. The software may reside entirely in the memory, or may also reside, completely or at least partially, within the RAM and/or within the processor during execution thereof by the computer system. Thus, the memory and the processor also constitute a system comprising machine-readable code.

In alternative embodiments, the machine operates as a standalone device or may be connected, e.g., networked to other machines, in a networked deployment, the machine may operate in the capacity of a server or a client machine in server-client network environment, or as a peer machine in a peer-to-peer or distributed network environment. The machine may be, for example, a computer, a server, a cluster of servers, a cluster of computers, a web appliance, a distributed computing environment, a cloud computing environment, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine. The term "machine" may also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

The foregoing disclosure of the exemplary embodiments of the present invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the forms disclosed. Many variations and modifications of the embodiments described herein will be apparent to one of ordinary skill in the art in light of the above disclosure. The scope of the invention is to be defined only by the claims appended hereto, and by their equivalents.

Further, in describing representative embodiments of the present invention, the specification may have presented the method and/or process of the present invention as a particular sequence of steps. However, to the extent that the method or process does not rely on the particular order of steps set forth herein, the method or process should not be limited to the particular sequence of steps described. As one of ordinary skill in the art would appreciate, other sequences of steps may be possible. Therefore, the particular order of the steps set forth in the specification should not be construed as limitations on the claims. In addition, the claims directed to the method and/or process of the present invention should not be limited to the performance of their steps in the order written, and one skilled in the art can readily appreciate that the sequences may be varied and still remain within the spirit and scope of the present invention.

What is claimed is:

1. A method comprising:
   receiving an item of content;
   parsing the item of content with a microprocessor to generate a linguistic annotated item of content with language associations;
   retrieving from a term selection rules repository stored upon a memory at least a rule of a plurality of rules;
applying with the microprocessor the at least a rule of the plurality of rules to establish a set of candidate sentiment carrying terms within the linguistic annotated item of content;

querying the set of candidate sentiment carrying terms against a target-domain sentiment lexicon to generate a set of sentiment labeled terms; and

applying to the linguistic annotated item of content a set of sentiment labeling rules established in dependence of at least the set of sentiment labeled terms to generate a sentiment label for the item of content.

2. The method according to claim 1 wherein, the language associations are at least one of parts of speech, phrasal elements, and grammatical relations associated with terms that form a predetermined portion of the item of content.

3. The method according to claim 1 wherein, each sentiment labeled term is associated with at least one of a sentiment label and a sentiment intensity.

4. The method according to claim 3 wherein, the at least one of the sentiment label and the sentiment intensity are employed in the application to the linguistic annotated item of content of the set of sentiment labeling rules.

5. A method comprising:

   a) receiving an item of content;

   b) receiving upon a microprocessor an indication of a predetermined portion of the item of content to analyze;

   c) establishing with the microprocessor a plurality of positive sentiment terms and a plurality of negative sentiment terms;

   d) parsing with the microprocessor the predetermined portion of the item of content to count occurrences of a negative sentiment term of the plurality of positive sentiment terms to establish a positive sentiment count;

   e) parsing with the microprocessor the predetermined portion of the item of content to count occurrences of a negative sentiment term of the plurality of negative sentiment terms to establish a negative sentiment count; and

   f) determining with the microprocessor a sentiment label to associate with the item of content in dependence upon at least one of the occurrences of the positive sentiment term and occurrences of the negative sentiment term.

6. The method according to claim 5 wherein, each positive sentiment term of the plurality of positive sentiment terms has an associated positive intensity level;

   each negative sentiment term of the plurality of negative sentiment terms has an associated negative intensity level.

7. The method according to claim 6 wherein, counting occurrences of the positive sentiment terms of the plurality of positive sentiment terms is achieved by:

   determining a number of occurrences for each positive sentiment term;

   multiplying the number of occurrences for each positive sentiment term by its respective intensity level to generate a weighted occurrence count;

   summing the resulting weighting occurrence counts for the plurality of positive sentiment counts to generate the positive sentiment count; and

   counting occurrences of the negative sentiment terms of the plurality of negative sentiment terms is achieved by:

   determining a number of occurrences for each negative sentiment term;

   multiplying the number of occurrences for each negative sentiment term by its respective intensity level to generate a weighted occurrence count;

   summing the resulting weighting occurrence counts for the plurality of negative sentiment counts to generate the negative sentiment count.

8. The method of claim 5 further comprising:

   establishing a number of predetermined portions of the item of content in step (a) and associating with each predetermined portion of the item of content a portion weighting;

   steps (b) to (e) are repeated for a number of predetermined portions of the item of content; and

   step (f) now comprises multiplying for each predetermined portion of the item of content the positive and negative sentiment counts by the respective portion weighting for that predetermined portion of the item of content to generate portion weighted positive and negative sentiment counts respectively and summing the results for all predetermined portions of the item of content.

9. The method according to claim 5 further comprising:

   determining with the microprocessor a domain associated with the item of content in step (a); and

   selecting with the microprocessor a sentiment lexicon of a plurality of sentiment lexicons, the selection made in dependence upon at least the domain.

10. The method according to claim 5 wherein, determining the sentiment label is at least one of:

    also dependent upon the imbalance between the counts of occurrences of the positive sentiment term and negative sentiment term; and

    selecting a sentiment label that is not one of either the positive sentiment term or negative sentiment term used in establishing the occurrences.

11. The method according to claim 5 wherein, generating the sentiment label is achieved in dependence upon at least one the difference, the sum, the ratio of the occurrences of the positive sentiment term and occurrences of the negative sentiment term, the positive sentiment term, and the negative sentiment term.

12. The method according to claim 5 wherein, generating a psychological tone qualification in dependence upon at least one the difference, the sum, the ratio of the occurrences of the positive sentiment term and occurrences of the negative sentiment term, the positive sentiment term, and the negative sentiment term.

13. The method of claim 5 further comprising:

    repeating step (d) for each positive sentiment term of the plurality of positive sentiment terms and each negative sentiment term of the plurality of negative sentiment terms; and

    step (f) now comprises summing the results for all of the plurality of positive sentiment steps (f) now comprises with the microprocessor the sentiment label to associate with the item of content in dependence upon at least one of the occurrences of all positive sentiment terms of the plurality of positive sentiment terms and occurrences of all negative sentiment terms of the plurality of the negative sentiment terms.

14. The method according to claim 11 further comprising:

    generating a psychological tone qualification in dependence upon at least one of the distribution of occurrences
of all positive sentiment terms of the plurality of positive sentiment terms and the distribution of occurrences of all negative sentiment terms of the plurality of negative sentiment terms.

15. The method according to claim 5 further comprising: determining with the microprocessor a domain associated with the item of content in step (a); and determining a sentiment to associate to an item of content, the determination being in dependence upon at least the domain and the sentiment label.

16. A method comprising:
   receiving with an item of content; 
   processing with a microprocessor the item of content to determine occurrences of content sentiment-carrying terms; 
   displaying to a user the sentiment labels of content sentiment-carrying terms within the item of content; and presenting to the user any sentiment intensity variation based on matching at least one of a predetermined sentence and a phrasal syntactic structure of the document with a repository of syntactic structure patterns.

17. The method according to claim 16 wherein, the sentiment intensity variation is at least one of an increase, a decrease, neutralization and a reversal.

18. The method of claim 16 wherein, 
   describing any sentiment intensity variation is based upon matching the sentiment of at least two adjacent sentiment-evaluated sentences with the repository of syntactic structure patterns.

19. The method of claim 16 further comprising, allowing the user to select at least one of the sentiment carrying terms, sentences and rhetorical structures to access an explanation relating to how the derived sentiment label is associated with the clicked entity.

20. A method comprising:
   a) receiving a plurality of items of content; 
   b) identifying with a microprocessor within the plurality of items of content at least a core multi-item concept of a plurality of core multi-item concepts, each core multi-item concept relating to a concept contained at least within a predetermined portion of the plurality of items of concept; 
   c) selecting a core multi-item concept from the plurality of core multi-item concepts; and 
   d) establishing with the microprocessor a sentiment relating to the core multi-item concept for the plurality of items of content.

21. The method according to claim 20 wherein, 
   the sentiment relating to the core multi-item concept for the plurality of content is established by at least one of: 
   e) determining a count based sentiment for the core multi-item concept for each item of content of the plurality of items of content; and 
   establishing the sentiment in dependence upon at least the plurality of document count based sentiment; and 
   f) determining a context count based sentiment by identifying each instance of the core multi-item concept within the plurality of items of content.

22. The method according to claim 20 further comprising: 
   repeating steps (c) and (d) for a predetermined subset of the plurality of multi-item concepts; and 
   presenting at least one of the predetermined subset of the plurality of multi-item concepts to the user together with its associated sentiment.

23. The method according to claim 20 further comprising:
   e) receiving a second plurality of items of content; 
   f) repeating steps (c) and (d) for the same core multi-item concept; 
   g) presenting to a user at least one of: 
   the original sentiment and a variance established in dependence upon at least the original sentiment and the new sentiment.

* * * * *