



- (51) **International Patent Classification:**
G06F 16/34 (2019.01)
- (21) **International Application Number:**
PCT/EP2022/064494
- (22) **International Filing Date:**
27 May 2022 (27.05.2022)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
21176102.8 27 May 2021 (27.05.2021) EP
- (71) **Applicant: PARLAMETRIC AB** [SE/SE]; Grönegatan 4A, 222 22 Lund (SE).
- (72) **Inventor: KÄLLSTRAND, Johan**; Vegagatan 1, 224 57 Lund (SE).
- (74) **Agent: KIPA AB**; Drottninggatan 11, 252 21 Helsingborg (SE).

MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— with international search report (Art. 21(3))

(81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA,

(54) **Title:** CHARACTERIZATION OF TEXT-BASED DATA

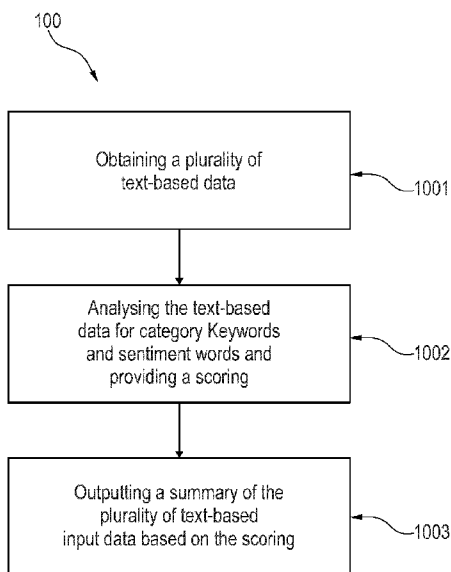


Fig. 1

(57) **Abstract:** The disclosure relates to a computer implemented method of characterizing text-based data. The method comprising obtaining a plurality of text-based data from a plurality of users or sources, wherein each text-based data of the plurality of text-based data is provided by a user or a source from the plurality of users or sources, the text-based data is unstructured; analysing each of the text-based data for at least one of category words and unique words, and related sentiment words and/or expressions surrounding the at least one of category words and unique words; providing a scoring of each text-based data based on the at least one of category words and unique words, and related sentiment words and/or expressions surrounding the at least one of category words and unique words; and outputting a summary of the plurality of text-based data based on the scoring.



SPECIFICATION**TITLE: CHARACTERIZATION OF TEXT-BASED DATA**5 **BACKGROUND OF THE INVENTION****Field of the Invention**

This disclosure pertains to characterizing text-based data, wherein the text-based data is unstructured. In particular, the disclosure relates to a computer
10 implemented invention wherein user entered text-based data is categorized and structured to extract relevant information.

15 **Description of the Prior Art**

Adaption to real-time insight from customers, markets and employees is one of the most important success factors for any company.

But organizations drown in data they can't use,
20 rather than benefit from it. Others collect thousands of valuable opinions and ideas without means to extract meaning from them.

Every day, billions of valuable customer and employee opinions rotten as time makes them obsolete. Most systems
25 for automatic narrative interpretation give the user more complex data instead of actionable insights. Surveys with specific questions or predefined scales and social media metrics counting likes and clicks are misleading and biased and does not reveal why someone clicked or liked content.

30 Another problem in automatic narrative analytics is when applying standard machine-learning algorithms to build dictionaries, huge amounts of training data are needed. This leads to narrative analytics being only possible to use for enormous data flows, not for niche areas.

35 A method enabling large sample faster, more in-depth open question where it is possible to extract more

information such as trends and insights would be advantageous. Especially if it makes it possible for real-time tracking of markets and customers in their interaction channels.

5

SUMMARY OF THE DISCLOSURE

Accordingly, embodiments of the present disclosure preferably seek to mitigate, alleviate, or eliminate one or more deficiencies, disadvantages or issues in the art, such as the above-identified, singly or in any combination by providing a method, such as a computer implemented method, a computer program and a data processing apparatus for characterizing text-based data. The disclosure makes it possible to go through narrative data, even in real-time, to automatically extract and generate trends and/or insight. The extracted and generated trends and insights in the data is based on analysis with statistical significance. The narrative data may be surveying sources, social media, customer interactions, digital media and/or traditional media. Because the input data is text-based and unstructured, such as story driven, and the analysis is performed using natural language processing algorithms, click and like metrics can be avoided and instead the output may provide analysis about why the users, customers etc likes or dislikes certain aspects.

In one aspect of the disclosure, a computer implemented method of characterizing text-based data is described. The method allows extraction of at least one characterizing sentiment information representative for a plurality of text-based data.

The method may include obtaining a dataset comprising the plurality of text-based data from a plurality of users or sources. Each text-based data of said plurality of text-based data may be provided by a user or a source from the

35

plurality of users or sources. The text-based data may be unstructured.

The method may further include analysing each of the text-based data for at least one of category words and
5 unique words, and related sentiment words and/or expressions surrounding the at least one of category words and unique words using a dictionary.

The method may also include providing a scoring for each text-based data by quantifying the at least one of
10 category words and unique words, and related sentiment words and/or expressions surrounding the at least one of category words and unique words.

The method may include obtaining statistical information of the dataset based on the scoring of each
15 text-based data.

The method may also include outputting a summary of comprising characterizing sentimental information representative for the dataset based on the statistical information.

20 By using expressions, whole or parts of phrases and sentences, as well as word stems, may be used in the analysis. The use of expressions together with labelling the context makes it possible to conduct complex sentiment analysis and find other meanings than by just looking at
25 category words and unique words, and related sentiment words.

This may make it possible to analyse smaller datasets than what may normally be done.

The method may also include outputting a summary of
30 the plurality of text-based data based on the scoring.

In a second aspect, a computer program is described. The computer program includes instructions which, when the program is executed by a computer, cause the computer to carry out the method here in described.

In a third aspect, a computer-readable medium is described. The computer-computer readable medium may include instructions which, when executed by a computer, cause the computer to carry out the herein described
5 method.

In a fourth aspect, a data processing apparatus is described. The data processing apparatus may include means for carrying out the method herein described.

10

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects, features, and advantages of which examples of the disclosure are capable of will be apparent and elucidated from the following description of
15 examples of the present disclosure, reference being made to the accompanying drawings, in which:

Fig. 1 is illustrating of a schematic example of the method;

Fig. 2 is illustrating a schematic diagram to
20 illustrate categories and sentiment words/expressions but also some basics of the scoring;

Fig. 3 is illustrating of a schematic example of a flow-chart related to analysis of text-based input data and generating a dictionary;

25 Fig. 4 is illustrating a schematic 2D-vector for performing the scoring;

Fig. 5 is illustrating a schematic example of a 2D-vector for performing the scoring;

30 Fig. 6 is illustrating an example of how output data may be summarized based on an obtained score;

Fig. 7 is illustrating a representation of a result including representative quotations for each cluster;

Fig. 8 is illustrating factors that explained >90% of the positive impressions found in the analysis;

Fig. 9 is illustrating representative quotes from the comments obtained from the analysis; and

Fig. 10 is illustrating a narrative analysis of completely open text responses.

5

DESCRIPTION OF EXAMPLES

Specific examples of the disclosure will now be described with reference to the accompanying drawings. This disclosure may, however, be embodied in many different forms and should not be construed as limited to the examples set forth herein; rather, these examples are provided so that this disclosure will be thorough and complete, and will fully convey the scope of the disclosure to those skilled in the art.

The following disclosure focuses on examples of the present disclosure applicable for characterizing text-based data, such as unstructured narrative data. The disclosure makes it possible to go through unstructured narrative data, even in real-time, to automatically extract and generate trends and/or insight from it. The extracted and generated trends and insights in the data is based on analysis with statistical significance. The text-based data may be survey sources, such as forms to fill out with unstructured text, social media, customer interactions, digital media and/or traditional media etc. Because the input data is text-based and unstructured, such as story driven, and the analysis is performed using natural language processing algorithms, click and like metrics can be avoided and instead the output may provide analysis about why the users, customers etc likes or dislikes certain aspects. The method may also be used to find variations, such as anomalies or trends, in a flow of text-based data.

Such a method will provide the advantage of adapting to real-time insights and/or trends from users, for example, customers, markets, politics, and/or employees. Examples could be to find out about what employees are
5 satisfied with or dissatisfied with at their workplace, what customers thinks about certain products, finding trends in social medias or in customer interactions. Trends could be related to political trends, such as geopolitical trends. Analysing text-based data, such as from online
10 media, social media and/or traditional media etc, may be used for predicting trends in the financial market based on hot topics and/or to catch new trends or changes that could affect the financial market and/or provide investment opportunities at an at an early stage.

15 The method disclosed herein may be implemented on a computer or a computer system, such as a computer network or a server. The method may be implemented by special-purpose software (or firmware) run on one or more general-purpose or special-purpose computing devices. In this
20 context, it is to be understood that each "element" or "means" of such a computing device refers to a conceptual equivalent of a method step; there is not always a one-to-one correspondence between elements/means and particular pieces of hardware or software routines. One piece of
25 hardware sometimes comprises different means/elements. For example, a processing unit serves as one element/means when executing one instruction but serves as another element/means when executing another instruction. In addition, one element/means may be implemented by one
30 instruction in some cases, but by a plurality of instructions in some other cases. Such a software-controlled computing device may include one or more processing units, e.g. a CPU ("Central Processing Unit"), a DSP ("Digital Signal Processor"), an ASIC ("Application-Specific Integrated Circuit"), discrete analog and/or
35

digital components, or some other programmable logical device, such as an FPGA ("Field Programmable Gate Array"). The data processing device 10 may further include a system memory and a system bus that couples various system

5 components including the system memory to the processing unit. The system bus may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The system memory may include computer

10 storage media in the form of volatile and/or non-volatile memory such as read only memory (ROM), random access memory (RAM) and flash memory. The special-purpose software may be stored in the system memory, or on other removable/non-removable volatile/non-volatile computer storage media

15 which is included in or accessible to the computing device, such as magnetic media, optical media, flash memory cards, digital tape, solid state RAM, solid state ROM, etc. The data processing device 10 may include one or more communication interfaces, such as a serial interface, a

20 parallel interface, a USB interface, a wireless interface, a network adapter, etc., as well as one or more data acquisition devices, such as an A/D converter. The special-purpose software may be provided to the control unit or data processing device on any suitable computer-readable

25 medium, including a record medium and a read-only memory.

Fig. 1 is illustrating of a schematic example of the claimed method 100, which may be a computer implemented method, of characterizing text-based data. The text-based data may be unstructured narrative text.

30 A first step includes obtaining a plurality of text-based data 1001. The text-based data may be obtained from a plurality of users and/or sources. For example, each text-based data of the plurality of text-based data may entered by a user from the plurality of users. In other examples,

35 the text-based data may be published information, such as

information published in social media, digital media and/or traditional media. The text-based data may be obtained from surveying sources, social media, customer interactions, digital media and/or traditional media. The text-based data
5 may be full narrative text, such as story-driven and not based on clicks and likes metrics. An example could be a reply to an open question, such as posted on a webpage and/or through social media.

The text-based data may be obtained from are a survey
10 made as a dedicated homepage with open questions to be answered by users. Additionally, and/or alternatively, the text-based data may also be obtained from comments left in social media, either in response to a dedicated question or spontaneous comments to an article, video clip, commercial
15 and/or in response to a comment by another user. Additionally, and/or alternatively, the text-based data may also be obtained from articles published in traditional media and/or digital media, such as blogs, bulletins or forums. Additionally, and/or alternatively, the text-based
20 data may be obtained by e-mail, SMS and/or other types of text-based communication means.

A second step may include analysing 1002 each of the text-based data for category words, unique words and sentiment words and providing a scoring of each text-based
25 input data. The scoring may be based on the occurrence of keywords related to the categories and/or unique words and the sentiment connected to these keywords and/or unique words. The sentiment may be sentiment words and/or expressions, but could also include symbols such as
30 exclamation marks or emojis, such as smiley symbols. The categories may include a plurality of keywords related to a particular category. Unique words are words that does not belong to a category or words which has not yet be categorised.

By using expressions, whole or parts of phrases and sentences, as well as word stems, may be used in the analysis. The use of expressions together with labelling the context makes it possible to conduct complex sentiment
5 analysis and find other meanings than by just looking at category words and unique words, and related sentiment words.

This may make it possible to analyse smaller datasets than what may normally be done.

10

The analysis may utilize a dictionary created for the category and sentiment words. The dictionary may also include unique words. The dictionary may be built using open source data and also trained on open-source data. The
15 dictionary may be obtained and trained by Artificial Intelligence (AI). In some examples are the dictionary for the category words, i.e. the keywords related to each category, and/or unique words created using AI, and the sentiment words may be a pre-defined list of sentiment
20 words and expression.

The analysis may be done by a Natural Language Processing algorithm. Preferably, an algorithm having content adaptive capability to be able to more accurately analysing different type of areas, such a niche area. This
25 may be done by automatically label context or themes for each dictionary. By automatically labelling context or themes, the analysis may be adapted depending on the dataset. The adaption may help to compensate for differences in language discourse. For example, words or
30 syntaxes could have different meaning depending on the area. For example, may identical words used in different areas, such as fashion customers, gamers, building, car owners, employees, and patients, have different meanings. This may be done by adapting the dictionary depending on
35 the area of text-based input area using auto-creation.

Adaption of the dictionary may also mean adaption of the training using the AI.

The analysis and scoring may be based on quantifying the category words and sentiment words surrounding keywords related to the categories. This may be done by using proximity, such as a moving window, to quantify the category and sentiment words surrounding a keyword. The size of the moving window may be obtained from the training of the dictionary. The moving window may in some examples change its size dynamically depending on the input data. In other examples may the size of the moving window is fixed to typical distances between category keywords or unique words and sentiment words. The size of the window may depend on various factors, such as the source of the text-based input data, the categories to be evaluated, etc. For example, depending on the source of the text-based input data, the sentiment level may need to be accounted for. The sentiment level may relate to how the sentiment words are used, for example, some type of text-based input data has a high level of sentiment words and expressions while other type of texts has a very low level. Depending on the source, the sentiment level may vary between users providing the text-based input data. This may be accounted for, at least to some extent, by the training the dictionary to obtain an optimal proximity window.

Additionally, and/or alternatively, other filters may be used either by itself and/or in combination with a proximity window. Examples of filters may include, exclude and/or rule-based filters with a variation of criterions. Other type of filters may be RegEx (Regular Expression Algorithms) and other types of propositional calculus or sentential logic.

In some examples, training may be used to predict the sentiment level of the text-based input data so that it can

be accounted for when performing the quantification and scoring.

Setting the proximity window and/or predicting the sentiment level, may be useful to give all text-based input data used in an analysis the same weight. When the text-based input data comes from various sources and/or users, differences in the sentiment level between various input data may otherwise give some sources and/or users having a high sentiment level a weight than sources, users having a lower sentiment level. By being able to level various sources and/or users, their information may be given the same weight in the analysis and in the final result.

A sentiment word or expression is a word or expression that may convey a positive or negative view. A sentiment word or expression could also give a neutral sentiment. Other types of emotions may be associated with the sentiment words, such as the five base emotions, e.g. joy, fear, sadness, disgust and anger. Depending on the word or expression used and its location in relation to a keyword related to a category or unique word, a scoring is allocated. For example, is the word or expression located before or after the keyword, and how many words are in-between the keyword and sentiment word or expression. The algorithm may also take into account how many category and/or sentiment words are associated to a keyword when providing a scoring for the text-based data.

The algorithm used for the analysis and the scoring may be based on a vector system. In some examples, the vector system is a 2-dimensional vector.

By adapting this approach, the analysis may be performed with statistical significance.

To further improve the analysis and provide more information, the algorithm may profile each of the text-based data provided as an input and clustering group of text-based data depending on determined profiles.

The method may also include outputting 1003 a summary of the plurality of text-based data based on the scoring. The summary may include statistic over the sentiment and the categories in the text-based data
5 provided by the users. The statistic may be clustered to provide information related to different aspects.

As part of the summary, the analysis may provide representative quotes based on the scoring. The quotes may come from the user input data or be synthetically
10 constructed. Different quotes may be used to for different sentiments or categories, but also related to different clusters.

Another possibility is to use text-based data from multiple sources, such as, for example, at least one of
15 social media, digital media, traditional media, and surveys, and performing cross-validation between the different sources.

Using multiple sources may not only provide more data but by using cross-validation between sources a better
20 representation of the reality may be obtained from the analysis. A better representation may provide a better accuracy in the output data. For example, different people use different medias and also express different types of opinions based on the media and the anonymity the media
25 provides. This may include different sentimental levels.

Cross-validation using multiple sources may therefore be used to validate the relevance of a result. Cross-validation between multiple sources may be used to remove
30 data that is not relevant. Data that may not be relevant could be social media phenomenon, such as trolls, miss leading information and false information.

Multiple-sources validation may be used to create better dictionaries for the narrative analytics to reach more valid and specific hits of categories and sentiments.

Fig. 2 is a schematic diagram 200 to illustrate categories and sentiment words/expressions but also some basics of the scoring. The schematic diagram 200 is only used to illustrate what the different terms are and how they are used. Along the y-axis, the categories 10 are illustrated. In the illustrated example, the categories are "SALARY" and "LEADERSHIP". In some examples, these categories may be sub-categories to a main category, such as "WORKPLACE".

10 In the illustrated example, one exemplary text-based input data is provided which has been divided into two parts 12a, 12b. In the illustrated example, the first part of the exemplary text-based input data reads "Too low salary" 12a, and the second part reads "management is
15 great" 12b. In the example, the words or expressions "low" and "great" are sentiments. "low" is defined as a negative sentiment 11a and "great" is defined as a positive sentiment 11b. A negative sentiment 11a is in this example provided with a negative scoring which is illustrated by
20 the positioning the first part of the exemplary input data 12a in the negative half of the x-axis. A positive sentiment 11b is in this example provided with a positive scoring which is illustrated by the positioning the second part of the exemplary input data 12b in the positive half
25 of the x-axis. Sentiment words and/or expressions having a neutral meaning can also be considered as part of the scoring.

The term "salary" in the first part of the exemplary input data 12a is a keyword connected to the category
30 "SALARY", and the term "management" in the second part of the exemplary input data 12b is a keyword connected to the category "LEADERSHIP". By going through the text-based input data, and quantifying the keywords related to each category and the associated sentiments words and/or

expressions surrounding each keyword, a scoring can be created for each category.

As previously described herein, further characteristics can be taken into account when performing the quantification and obtaining the scoring. Examples of characteristics that may be included are positioning of the sentiment word and/or expression, such as before or after the keyword; how many sentiment words and expressions are used for each keyword; how many keywords associated by a category is provided in each text-based input data; the distance between the keyword and the connected sentiment word and/or expression, etc. By profiling each of the text-based data provided as an input and clustering group of text-based data depending on determined profiles, it may be possible to provide further information from the input data, such as differences between users and/or sources which are to some extent connected. For example, profiling and clustering may be used to segmenting the output.

Fig. 3 is illustrating of a schematic example of a flow-chart 300 related to analysis of text-based input data and generating a dictionary. The method may include the following steps:

An initial step 2001 where a client has a question. Typically, the client wants an up to date understanding and analysis of open-text based question. The information to be analyzed may be obtained from scanning sources, such as social media, blogs, publications etc. The type of sources may be defined to be relevant for the analysis, such as within a particular area like finance, politics, marketing impact, product understanding etc. The information to be analyzed may also come from a company employment survey or analyzing a customer support speech. Other sources for the information to be analysed could be media publications, and so on.

The incoming words in the text-based input data may then be quantified 2002. This can be done by keeping identities of respondents to ensure one input per identity but at the same time making the outcome completely
5 anonymous.

To clear up the data, all non-meaning bearing words may be filtered out 2003. Non-meaning bearing words are typically words like "and", "or" and "the" and similar.

An auto-categorization and database check 2004 may
10 then be performed. This step is to ensure correct categorization. In this step, the remaining words from the step 2003 are checked to a database to quantify most occurring categories, for example by quantifying category related keywords in the text-based input data. The step may
15 include a validation 2004a with text-based data, such as open source data. The validation 2004a, may be defined as a category generation, or as an auto-creation. The category generation may be performed to improve the accuracy of the scoring. The category generation may be performed on
20 general open source data. During this step 2004a, the words from step 2003 may be checked with millions of open sources, such as social media, articles, and webpages. If a certain word occurs significantly more in one type of sources and not in others, the theme of those sources may
25 thus be the auto-generated category, or auto-creation, assigned to that word.

An alternative to perform the auto-generation or auto-creation on general open source data could be to use specific open sources. An advantage with this may be that
30 the databases created may be niched based on the input data and therefore more relevant and representative for the analysis, for example when an analysis is carried out for narrow subject-areas or topics. This may be performed by analyzing the text-based input data for sub-categories. The
35 sub-categories may then be linked to a main category. When

performing the auto-generated category, or auto-creation, instead on performing the validation of general open source data, the validation may be performed on opensource data within the area of the main category. The step may then be
5 performed as for the general open source data, by quantifying the words in the input data and compare them to the open source data within the main category. If a certain word occurs significantly more in one type of sources and not in others within the area of the main category, the
10 theme of those sources may thus be the auto-generated category, or auto-creation, assigned to that word. Thereby generating categories and sub-categories. The new category and sub-category connectors found through the open-source scanning may increase the correctness and specificity of
15 the analysis, in particular for niched fields.

Another step may be to setting proximity windows of sentiments 2005. To find proper sentiments (for example negative, neutral, positive), the area adequate rules of sentiment words occurrence in relation to the category
20 words may need to be found. This may be done by tracking open sources and measuring typical distances of sentimental words and/or expressions to category related words. This may provide a dynamic size of the window that could vary between different category words, sentimental words and/or
25 expressions, or characteristics of the text-based input data types etc. Alternatively, the proximity window is set to a pre-defined fixed distance.

Identifying sentiments 2006. The sentiments may be found in pre-defined lists or be generated using AI. All
30 sentiments may then be identified in the text-based input data.

Categorizing 2007 all non-sentiment and non-category words as being unique words. The unique words are the words that has not been categorized in step 2004, 2004a or

identified as sentiment words and/or expressions under step 2006.

When all the words have been categorized, either as a category word, as a unique word or as a sentiment word
5 and/or expression, the window sizes may be used 2008 to find occurrence of sentiments close to both categories and unique words.

Words categorized as unique may be further analyzed. The analysis may include, some but not all of the following
10 steps; finding categories 2011 within the defined context-specific proximity window defined in step 2005 for the unique words. Categories surrounding unique word may be quantified. Finding sentiments 2012 within the defined context-specific proximity window defined in step 2005 for
15 the unique words. Sentiments surrounding unique word are quantified. Finding sentiments 2013 within the defined context-specific proximity window defined in step 2005 for unique words' different categories found in step 2011. Also, the analysis may include finding categories 2014
20 within the defined context-specific proximity window defined in step 2005 for Unique words' different sentiments from step 2012.

The below is an example to try to further explain the auto-categorization, or auto-creation process. The input
25 data may come from a completely free text input to be filled in from any User Interface, could be spoken and converted to text, related to a free, non-leading question / topic.

30	From word lists, words are quantified		
	Boss	2	L
	Leader	3	L
	Salary	5	
	Jump	4	
35	King	4	

Cheese	3		F
Management	2	L	
Food	1		F
P2300	2		U

5

When scanning the words in other sources (the auto-categorization), the categories can be defined. Looking in public PDFs, the words marked L(=leadership) above were found mainly in Management PDFs. Hence, building a category validation of those words leads to an auto-generated categorization.

10

The words labelled F(=food) were mainly found in PDF resources tagged Food.

15

The PDF resources may not be restricted to PDFs but also contains online information like newspapers and webpages.

The advantage here is that the categorization can be dynamic and time-relevant, meaning that the categories does change over time and new ones are put forth (like crypto).

20

When the auto-categorization has taken place, the sentiment proximity window may be added to the equation, the occurrence of pre-defined negative or positive words, sad or angry ones and so on, before or after the category words defined above are identified. This can connect Leadership to Negativity, thus identifying that in a typical dataset there are specific management issues.

25

The windows may be category dependent and these windows may be set using machine-learning of finding the typical absolute distances between category building words and sentiment words, making a more valid analytic framework.

30

The words at some frequencies (occurrence), not being normal non-informational words (like "if" "but" "and" and so on) and not being category words are auto-identified

being Unique (U above) and these are also sentiment analyzed using the auto-set window mentioned above.

It is then possible to go back to previous steps - for example finding sentiments surrounding a unique word and also tracking categories surrounding a Unique word to
5 understand the unique word.

Fig. 4 illustrates a schematic 2D-vector 400 for performing the scoring, and Fig. 5 is illustrating a schematic example of a 2D-vector 500 for performing the
10 scoring. In the example, words related to the category "Leadership" has been found 8 times in the text-based input data, 5 times has the category word "Leadership" been used with a negative sentiment word and 3 times with a positive. The score is therefore -2 and the percentage, which
15 provides the overall sentiment for this category word is therefore -20%. Words related to the category "Salary" has been found 4 times, 2 times with a negative sentiment and 2 times with a positive. Hence, the overall sentiment may be considered as neutral.

Words related to the category "organization" has been
20 found 4 times, 1 time with a negative sentiment and 3 times with a positive. Hence, the overall sentiment may be considered as positive at 50%.

Fig. 6 is illustrating an example of how output data
25 may be summarized 600 based on an obtained quantification and scoring of unstructured text-based input date.

Example 1

30 The first example relates to computer game optimization. The herein described method was applied on 37456 Youtube comments written as a response to a published trailer for a new game. A software implementing the method categorized and structured data to find all topics and what
35 was good and bad.

The results were benchmarked to comparison games to find out whether the game would be a commercial success or not.

5 Out of the 37456 comments, 6494 comments were found to be clearly positive and 1741 comments were found to be clearly negative.

The most common factors behind the positive comments were the combination of a unique story/plot, characters and most importantly overwhelmingly graphics and gameplay. Reference to other titles, Dishonored and Bioshock, tended to be positive. Negative explained by Gameplay not to be on par with the trailer and to complex story. Fig. 7 illustrates a representation of the result including representative quotations for each cluster. Fig. 8 illustrates factors that explained >90% of the positive impressions. Box sizes are in proportion to the number of comments. As can be seen in the figure, graphics was extracted to be the most commentated category with highest positive sentiment followed by gameplay, story/plot, characters and music, sound, narration. Slightly less positive, but still with positive sentiment were references and comparisons to Bioshock and Dishonored.

25 Other statistical finding with representative quotes from the comments can be seen in Fig. 9.

25

Example 2

This survey related to Corona attitudes in four different countries. The number of participants were 2006 respondents from UK, USA, Italy and Germany.

30 In this example, the answers have been profiled and clustered into their geographical location. In Fig. 10, a narrative analysis of completely open text responses is presented. The herein described algorithm found 11 distinct categories whereof these are the five most common.

The present invention has been described above with reference to specific examples. However, other examples than the above described are equally possible within the scope of the disclosure. Different method steps than those
5 described above, performing the method by hardware or software, may be provided within the scope of the invention. The different features and steps of the invention may be combined in other combinations than those described. The scope of the disclosure is only limited by
10 the appended patent claims.

The indefinite articles "a" and "an," as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to mean "at least one." The phrase "and/or," as used herein in the
15 specification and in the claims, should be understood to mean "either or both" of the elements so conjoined, i.e., elements that are conjunctively present in some cases and disjunctively present in other cases.

20

CLAIMS

1. A computer implemented method of extracting at least
5 one characterizing sentiment information
representative for a plurality of text-based data,
the method comprising:

10 obtaining a dataset comprising said plurality of
text-based data from a plurality of users or
sources, wherein each text-based data of said
plurality of text-based data is provided by a user
or a source from said plurality of users or sources,
said text-based data is unstructured;

15 analysing each of said text-based data for at least
one of category words and unique words, and related
sentiment words and/or expressions surrounding said
at least one of category words and unique words
20 using a dictionary;

25 providing a scoring for each text-based data by
quantifying said at least one of category words and
unique words, and related sentiment words and/or
expressions surrounding said at least one of
category words and unique words;

30 obtaining statistical information of said dataset
based on said scoring of each text-based data; and

outputting a summary of comprising characterizing
sentimental information representative for said
dataset based on said statistical information.

2. The computer implemented method of claim 1, wherein said text-based data is a reply to an open question by said user.
- 5 3. The computer implemented method of any of claims 1 or 2, wherein said summary includes typical quotes representing said plurality of text-based data.
- 10 4. The computer implemented method of any of claims 1 to 3, analysing each of said text-based data using a vector system.
- 15 5. The computer implemented method of claim 4, wherein said vector is a 2-dimensional vector.
- 20 6. The computer implemented method of any of claims 1 to 5, wherein said analysis and scoring is based on quantify said at least one of category words and unique words, and related sentiment words and/or expressions surrounding said at least one of category words and unique words.
- 25 7. The computer implemented method of claim 6, using proximity, such as a moving window, to quantify said at least one of category words and unique words, and related sentiment words and/or expressions surrounding said at least one of category words and unique words.
- 30 8. The computer implemented method of any of claims 1 to 7, profiling each of said text-based data and clustering groups of text-based data depending on a determined profile.
- 35 9. The computer implemented method of any of claims 1 to 8, wherein said dictionary is created for said at least one of category words and unique words, and

related sentiment words and/or expressions surrounding said at least one of category words and unique words.

5 10. The computer implemented method of claim 9, wherein said dictionary is built using open source data.

10 11. The computer implemented method of any of claims 9 to 10, adapting said dictionary depending on the area of text-based data using auto-creation.

15 12. The computer implemented method of any of claims 1 to 11, wherein said a plurality of text-based input data is obtained from multiple sources, such as a source being at least one of social media, digital media, traditional media, and surveys, performing cross-validation between different sources.

20 13. A computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out the method of any of claims 1 to 12.

25 14. A computer-readable medium comprising instruction which, when executed by a computer, cause the computer to carry out the method of any if claims 1 to 12.

30 15. A data processing apparatus comprising means for carrying out the method of any of claims 1 to 12.

35

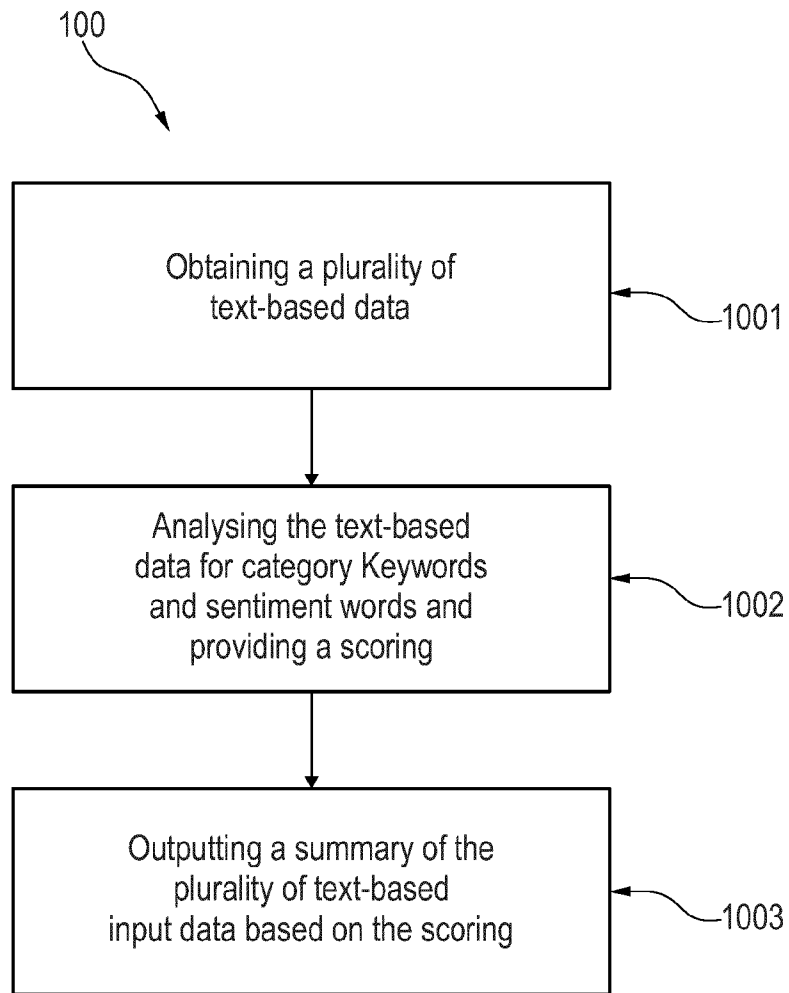


Fig. 1

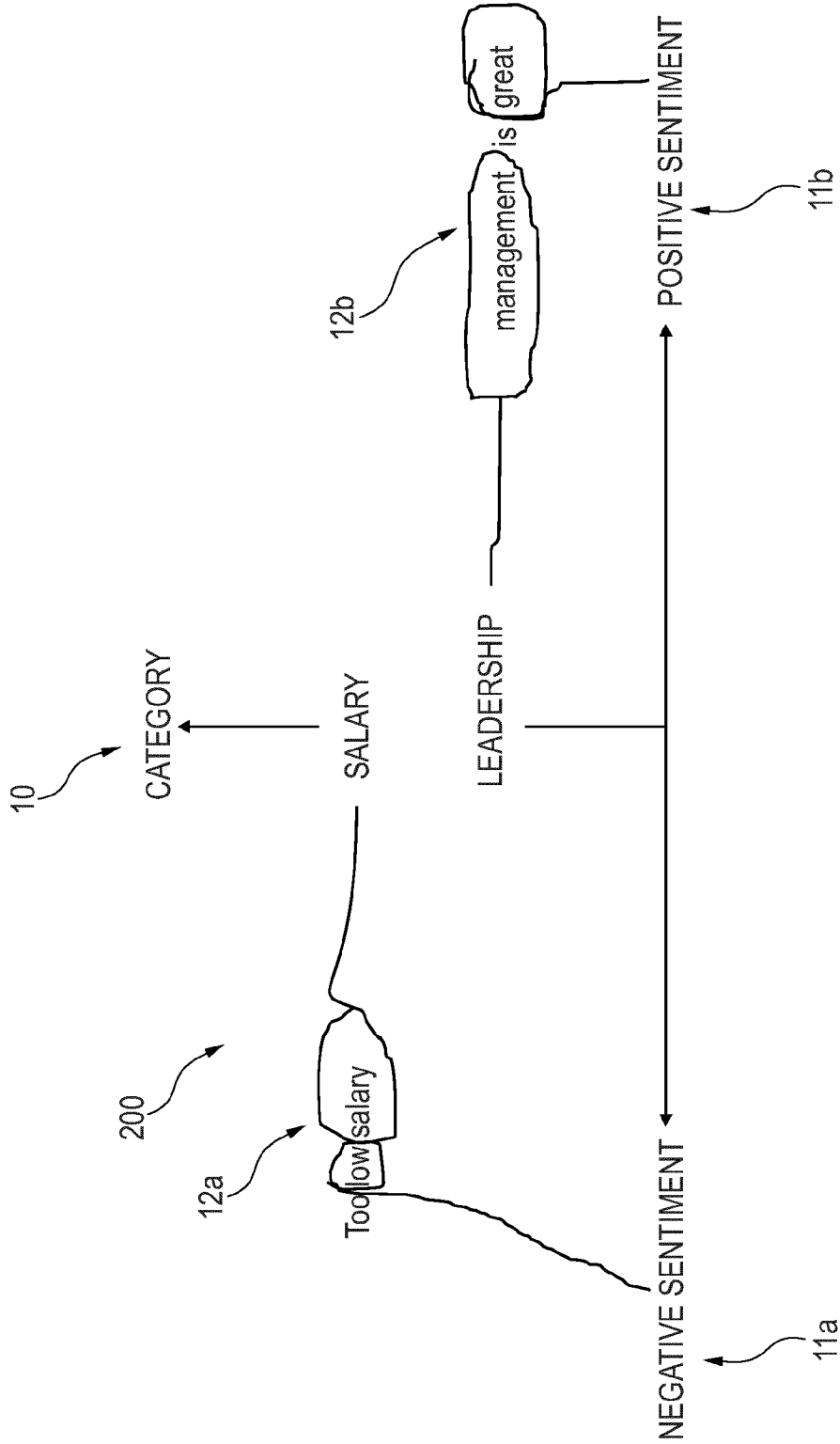


Fig. 2

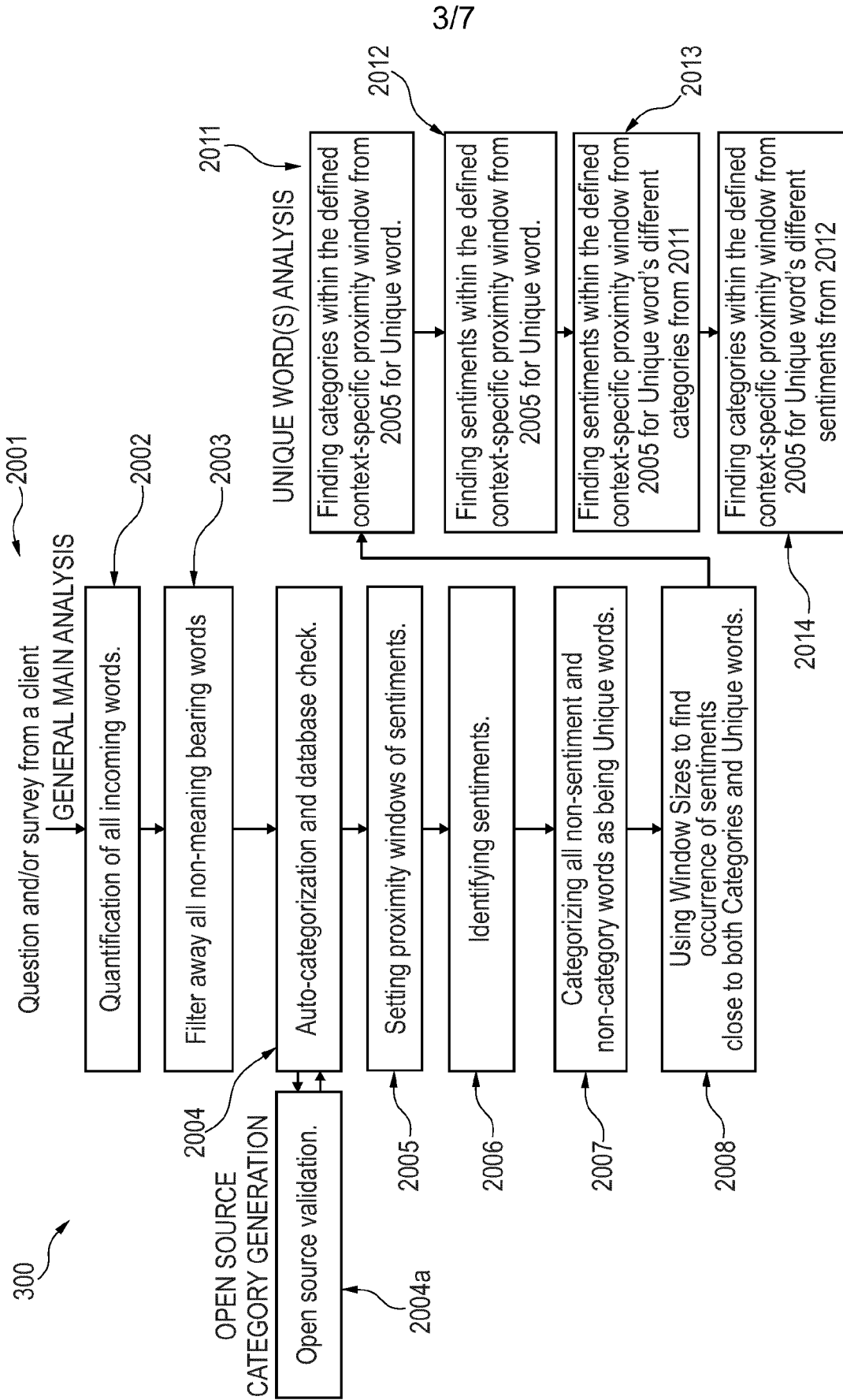


Fig. 3

4/7

400

Category words	Occurrence	Sentiment words						Count	Proc (%)
↓	↓	↓		↓	↓	↓	↓	↓	

Fig. 4

500

Category words	Occurrence	Bad	Sucks	Good	Great	Score	Proc (%)
Leadership	8	3	2	3	0	-2	-20%
Salary	4	1	1	0	2	0	0%
organization	4	1	0	2	1	2	50%

Fig. 5

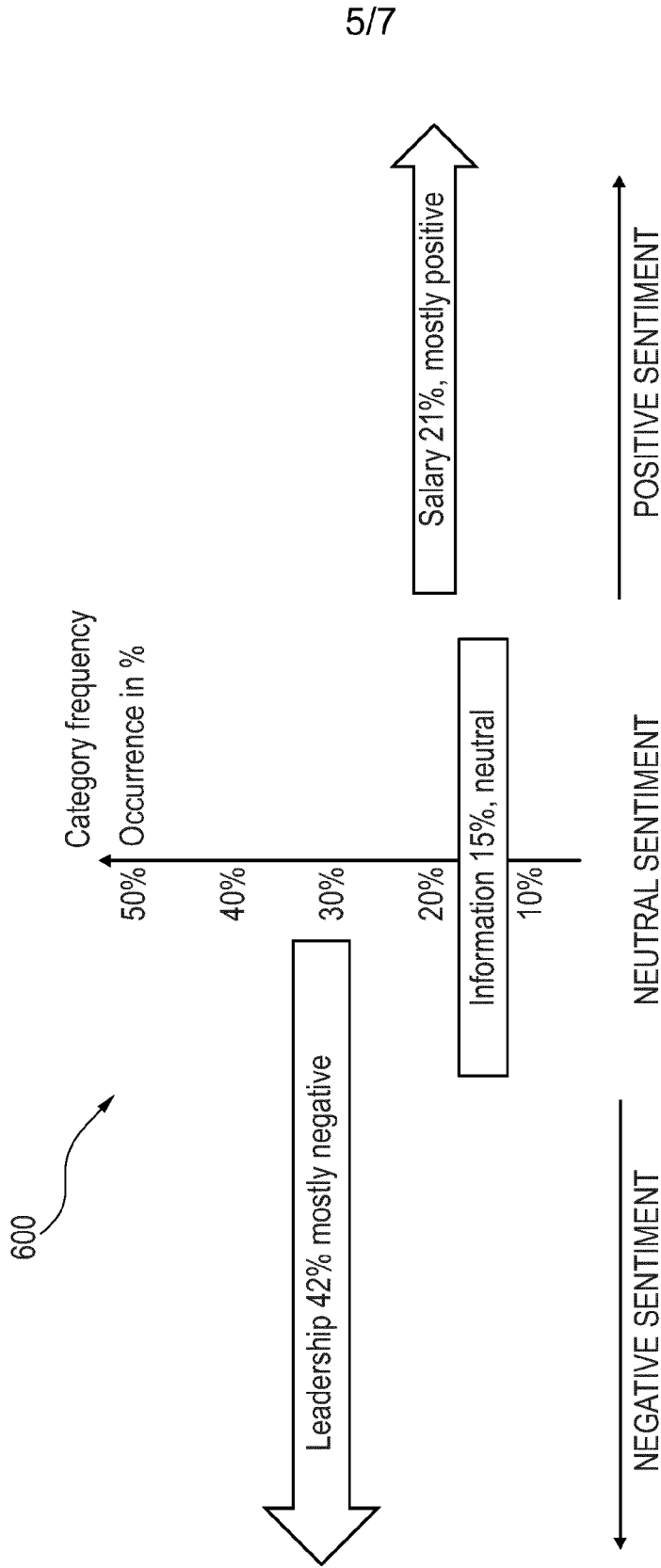
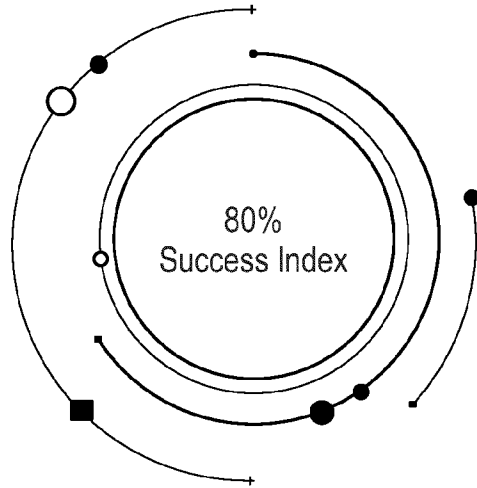


Fig. 6

6/7



General Impression / Graphics
 "i love the graphics and animations in this game"
 "this feels like bioshock 1 on acid. can't wait to play it!"
 "an indie game which was made by just one or a few people has better graphics than a game made by a huge company with millions of dollars."
 Explains **26%** of the positive comments

Gameplay
 "it is very interesting gameplay"
 Explains **21%** of the positive comments
 "the first 15 minutes was amazing and the idea is great but the gameplay fell flat on the demo" **Part of Negative cluster**

Complex Story
 "i am so confused with the storyline"
 "i don't get this game"

Story / Plot
 "i really love games with a deeper meaning. they let you be more creative in how you think, and you can decide what the story really is by yourself!"
 "it's amazing how beautiful and eerily rotten this world"
 Explains **11%** of the positive comments

Fig. 7

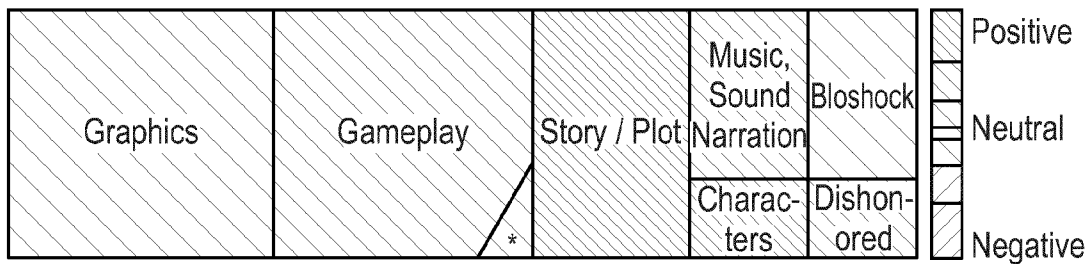


Fig. 8

7/7

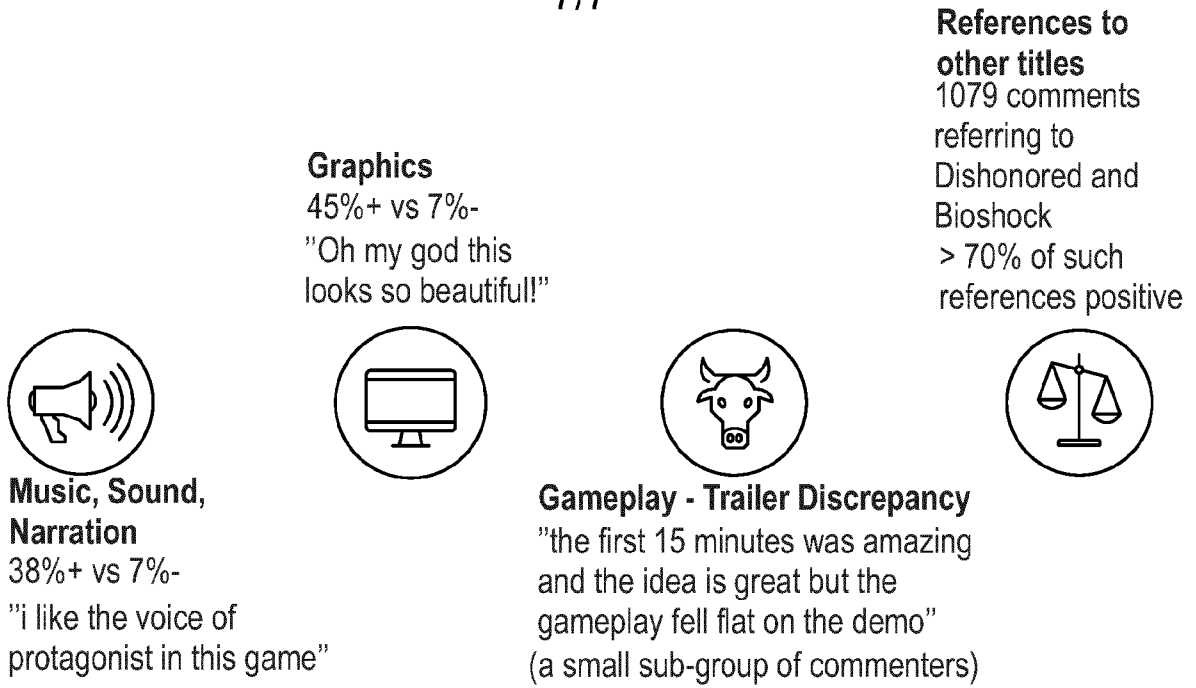


Fig. 9

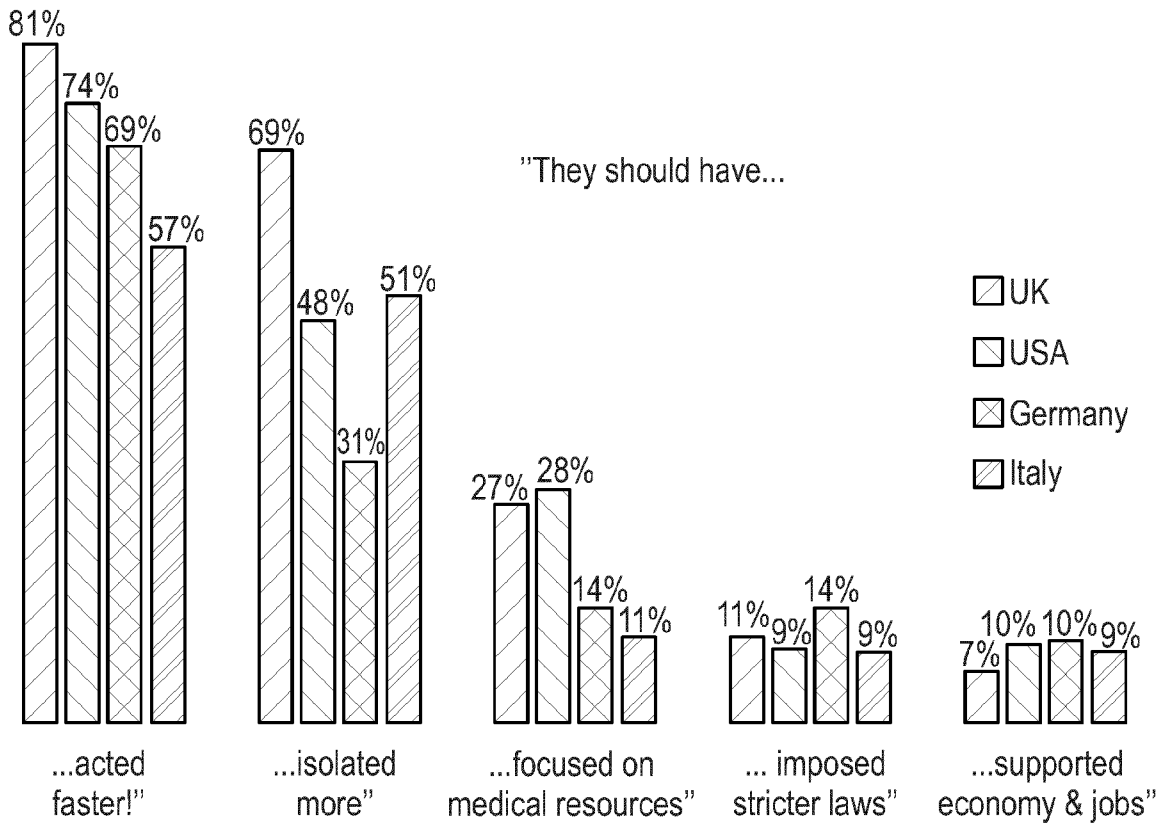


Fig. 10

INTERNATIONAL SEARCH REPORT

International application No PCT/EP2022/064494
--

A. CLASSIFICATION OF SUBJECT MATTER INV. G06F16/34 ADD. According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal, WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2020/279017 A1 (NORTON R DAVID [US] ET AL) 3 September 2020 (2020-09-03) abstract paragraphs [0005] - [0007], [0024] -----	1-15
X	US 2020/285662 A1 (CHATTERJEE ARINDAM [IN] ET AL) 10 September 2020 (2020-09-10) abstract paragraphs [0001], [0007] - [0009], [0022] - [0024]; figure 1 -----	1-15
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input checked="" type="checkbox"/> See patent family annex.		
* Special categories of cited documents :		
"A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family	
Date of the actual completion of the international search	Date of mailing of the international search report	
5 September 2022	13/09/2022	
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Authorized officer Siódmok, Wojciech	

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2022/064494

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2020279017 A1	03-09-2020	US 2020279017 A1	03-09-2020
		US 2022156464 A1	19-05-2022

US 2020285662 A1	10-09-2020	NONE	
