

A new Approach to Happiness Economics

Rajib VERMA
EHESS
rajibverma@yahoo.com

Abstract

Mining has been embraced by disciplines outside of computer science for over a decade. Yet, Happiness economics has not done so even though it has much to gain. An exposition of the technical, linguistic, and econometric difficulties involved in extracting happiness is combined with some solutions and a possible approach to mining. Possible research directions are discussed and the benefits that can be had are explored through the discussion of how data mining can illuminate questions relating to shocks, social contagion, social influence, and happiness.

Introduction

The increased use of the Internet by individuals, organizations, and media has created a reservoir of untapped data; especially since the popularization of online forums, discussion groups, and review sites. Information in the form of professional and amateur articles, blogs, chat sessions, message boards, technical papers, and reports appear on the net continuously and often faster than through traditional media. Some of it is factual, some is questioning, some is argumentative, and some of it is deeply opinionated—which suggests that hidden inside it is the way the author or community feels. If you could amass this information it could give you a real time indication of happiness, which would not only let you assess the current level of wellbeing, it would give you a time series showing the history, direction, magnitude of changes, and the impact of events like policy changes on happiness.

Text mining offers a way to do this by extending scientific knowledge through inductive data collection and

analysis, so by its very nature it tends to be empirical and atheoretical. The advantage is that, by letting the data speak for themselves, it has the ability to uncover unthought-of relationships. This is a little more difficult with conventional econometric methods since the data is usually compared to a model, which necessarily restricts you to a predefined paradigm. Pattern recognition on the other hand does not have this limitation, so any new and novel relations can be used to corroborate existing theories, or they can lead to new conjectures, and possibly, a paradigm shift. Unlike the usual serendipitous advancement of science, like the chance discovery of penicillin, it embodies a sound methodology and evolving technology that is capable of routinely discovering the unknown.

The mining process begins by defining the domain of study and then collecting data about it, either through some automated means (like a web crawler) or through an existing dataset. This collection is then cleaned, and the data is reduced to the relevant portions so that analysis can be done a little faster and easier. At this stage, a

method of analysis is chosen and executed, usually involving an algorithm that can deal with large and fragmented datasets consisting of a large number of variables. This makes it quicker and easier to find interesting patterns, which can then be further analyzed using econometric techniques if the necessary. Of course, the rest of the process is the same as in any other scientific approach: sound results are interpreted, and existing theories are supported, modified, or displaced.

Contrary to what is commonly believed, data mining is not just an assemblage of ad hoc correlations. There is a sound methodology and toolset that leads to new knowledge representations, and more importantly, there are several ways of evaluating the merit and the validity of the output. Merit refers to the qualitative features of the solution that make it worthy of consideration and it asks questions like: Is the result useful in relation to the research goals? Is it qualitatively different from what we knew before? Does it give greater depth of understanding or greater generality? Or, in spite of weaknesses, does it have redeeming qualities like providing better intuition or easier formulas? Validity on the other hand focuses on the correctness of the data and results. While there are some serious data collection and econometric issues (such as: the data being contaminated, it not being independent and identically distributed (IID), population shifts leading to non-stationarity, selection bias in the algorithms, non-randomness, and of course the usual measurement errors), there are ways to deal with these as discussed in detail below.

In general, though, the results can be subjected to tough statistical and economic significance tests to see if

they are drastically different from the null hypothesis. Since there is a wide collection of research using regular datasets the results can be compared to these, similar findings would lend credibility to the quality of the data and the methods used. In other cases, samples that give the same results but have been taken at different times, may also lend support to the validity of the findings.

While this type of research has not been done extensively for happiness studies yet, there is reason to believe that it could be a promising area since similar work has been done in finance, politics, and other fields with positive results. Zaima & Harjoto and Antweiler & Frank show that online discussion can provide better forecasts than financial analysts, Wysocki demonstrates that the intensity in volume of after hours postings can predict the variation in stock market returns, and that the higher the amount of discussion the higher the volatility. Das and Chen show that, in the presence of high ambiguity, time series and cross sectional aggregation of online data improves the quality of the polarity (i.e., sentiment) index.

These results suggest two things. First, subjective Internet discussions encapsulate all available information (much like prices do in the Hayekian / Austrian view), which leads to tighter predictions. This is consistent with the shortening of a distribution's variance as "n" (the number of data points) increases in conventional statistics. That is to say, we should find better predictability in discussions with more information (e.g. forums with more discussants or more diverse content). This provides a new venue for testing long held beliefs in economics, such as the idea that agents are efficiently able to make complex decisions using

rudimentary methods and available knowledge. Second, there seems to be a connection between Internet discussions and market activity. This could mean that discussions perform behaviour, or that online forums act as a sink for information and potential actions. In any event, they seem to represent something real, so they should be studied in detail.

This approach could easily be extended to all documents, not just the ones on the Internet. Data collected through ethnographic methods like interviewing could be used as an input to the crawler. Typically, qualitative researchers analyze text by finding categories that encapsulate the information inside. This approach has four main limitations that can all be addressed through mining: this classification scheme is time consuming, unique to each analysis, the results cannot be extended to the population, and it is difficult to make quantitative predictions.

The classifier could be designed to automatically search for existing categories, it could be programmed to find new ones using adaptive techniques, or as Camillo et. al. do, it could even use more traditional statistical data classification methods like factor analysis. Automating the process would not only speed up the analysis, it would provide greater insight into the variability of the underlying structures, and give an idea about the categories' degree of centrality. Something that would otherwise take many teams of independent researchers to do. By automating the process, it will be possible to analyze a large number of texts, which, if selected correctly, would make it possible to build representative models that could be extended the population at large.

Using a sample of 1200 happiness stories, Camillo et. al. create a discriminant statistical model that predicts what category the textual data belongs to. The confusion matrix shows promising results ranging from successful categorization rates of 42 percent to 100 percent.

Given that there are promising results from other fields, and that the underlying methodology is sound, there is an opportunity to extend it to the study of wellbeing. The remainder of the paper discusses several technical aspects of happiness mining in order to develop a framework to collect data, analyze it, and verify the results. It begins by discussing what makes happiness extraction difficult from technical, linguistic, and econometric perspectives; all the while posing solutions to them. This leads into an explication of how the extraction process can be implemented and a discussion of various classification techniques along with their pros and cons. The final section talks about the relation between happiness, social contagion, influence, and how these interactions can be measured using mined data.

Difficulty of extracting happiness

Technical and Linguistic Issues

The extension of HTML to XML has made parsing web pages easier since special tags have been included to mark particular pieces of information, like prices and product types. This has made some applications, like text mining for price comparisons, easier. However, sentiment extraction remains a more difficult issue because parsing, categorising and weighing online

content is inherently more difficult. Even though the W3C (the world wide web standards organization) has clearly indicated that the future of the web will be based on XML, and even though new tags could easily be developed to indicate things like the type and intensity of sentiment, it is not clear if this would make sentiment extraction easier since the author would have to encode his document manually. In some sense, this approach has the advantage that adding the tags directly imbeds the author's subjective opinions and sentiments directly into the document; but, realistically, only large publishing houses or professionals might do this, individuals would likely not. So methods need to be developed that can independently aggregate feelings, intensities, impact factors, and other indicators across the whole population. Otherwise, an enormous amount of information will be lost and results would likely be skewed.

Unlike commercial agents, individuals do not have an incentive to tag their text, so the meaning and degree of emotion has to be deciphered in other ways. This involves going through the content in order to find the subject of interest and the sentiment words that are associated with it. While there is a layering of information, one should keep in mind that human thought and behavior is much less structured than the data on e-commerce sites. In usual (simple) content parsing applications, there is a single layer of information, the object itself. In more sophisticated applications (like determining investor sentiment for particular stock), there are two layers of information, the writer and the object. Successfully measuring happiness involves navigating through the most layers: the writer, the various objects, and their association with happiness.

This layering of information makes it more difficult to assign the correct sentiment with the object and agent, and it makes it difficult to amalgamate everything into a single metric. Since it is possible for an agent to contribute to many sites under many pseudonyms, there is a possibility of multiple counting. Ordinarily this is a problem, but it can also be a boon since it may reduce the number of less informative equilibria¹, and because it indicates the degree of passion—so it may reflect the intensity of happiness in the network. In any event, the amount of over counting should be naturally limited since users tend to frequent the same sites.

Information layering is not the only thing that makes it difficult to assess, classify, and quantify sentiment. Social media (which includes wikis, blogs, forums, discussion and sharing sites) can often be ambiguous and composed of messages without any sentiment content (e.g., questions, factual responses, non-emotive or irrelevant opinion), as Admati and Pfleiderer show. Within eight hours of corporate press releases, on topic posts account for 50 to 75 percent of total postings. Of the on topic posts, between two and nine percent are questions, one to thirteen percent are factual, and between 40 and 65 percent are opinionated. This suggests that, at most, you can expect to find relevant sentiments in 49 percent of the messages, of course, this is an upper bound, and the useful information should be far less.

¹ Admati and Pfleiderer build a model showing that while sending exaggerated opinions reduces informativeness, in some cases it might eliminate less informative equilibria when broadcasting your opinion is costless. If it is costly, they show that overconfidence can improve informativeness.

Given that the sentiment we are interested in is unlikely to be tagged, uncovering it will have to be done manually or an effective way of automating the process will have to be developed. Even if you were to find a fantastic extractor, there would still be the issue of “aboutness” since many writings are unclear, even to a human. Pang, Lee, and Vaithyanathan test some supervised machine-learning methods and compare the results to a manual tagging method. They conclude that the automated methods are better, but that they are still far from being good.

The situation is complicated by linguistic structures. If you parse a text using its syntax, it becomes difficult to make inter-sentential connections and meaning is lost. This is the case when related subjects, sentiments, and strengths are separated by periods. In this case, each part forms a separate sentence and it becomes difficult for the algorithm to determine if they are connected, even though a human could do it easily. Take as an example “Russia is now run by the horrible Oligarchs. Ever since the fall of the union things have gotten far worse. This has made me and my family much more glum”. There are two subjects (Russia and Oligarchs) across three sentences, two sentiment words (worse and glum), and there are three quantification terms (horrible, far, and much more). An algorithm relying on syntax alone may not be able to quantify the sentences and assign them to the subjects Russia and oligarch even though they are semantically connected. There are ways to minimize this effect, for example by analyzing the contents inside a moving window; but, if the window is too small you will have the same problem,

and if the window is too large you will decrease performance.

The fact that the text can contain several parts complicates the evaluation. There could be many topics, some relating to happiness and others not. The challenge is not only to separate them into these groups, but also to find a way to keep track of sentiments relating to different topics and then combining them into an overall sentiment score. By doing this, you could develop sub-indexes and use them to enrich our understanding of what contributes to overall happiness and in what proportion. The reality is that doing these things is very difficult and has been comparatively unsuccessful in other applications. The reason is that most statistical analyzers either assume that they know what the topic is (which requires a priori knowledge or a genre classifier, which has its own technical limitations), or they look for key words in the text and assume that the sentiment is associated with them. The latter approach might work if there is only one subject, but when multiple key words are found or when the sentiments refer to different topics classification errors tend to emerge.

When doing cross-national comparisons, linguistic issues become more prevalent. Mining algorithms written for one language will have to be altered to accommodate the grammar, vocabulary, and expression differences in another. Proper comparison requires that mined data be similar, so care must be taken to ensure that the crawler selects the relevant sentiments. While some effort is needed to make the changes, this would be much simpler and cheaper than conducting surveys around the globe. Happiness mining could quickly yield large datasets and results

for every language group on the Internet. These could then be verified through comparison with existing datasets and results. Therefore, this approach could give us another way of estimating the distribution of happiness across the globe, provide a way of monitoring its changes on an ongoing basis, and it could provide a means of collecting and assessing the determinants of happiness for each linguistic group.

Aside from written media, sentiment is found in sources like video clips and audio files. Here the classification and quantification issues are even more complex since there is no easy way to parse the semantics. It could be done manually but that would be more difficult than with text since this type of media is much slower, and because careful attention would have to be paid to things like intonation, body gestures, and other forms of nonverbal communication. This would take an enormous effort and be very expensive.

Mining and Econometrics

Happiness studies is an emerging field that has developed mainly through surveys, econometric analysis, and experiments. Developing methods for collecting and analyzing sentiment content spread over networks would add an important tool to the existing arsenal of techniques, which would allow us to validate existing research using another approach. Similar findings would lend additional credibility to the work that has already been done, and any divergences would highlight those areas that need further study. At the same time, comparing the mining results to existing research would be a way to validate the quality of the extracted micro-level data.

Even though you can do exploratory analysis on existing datasets, there are

very real restrictions in the sense that the data is usually collected with certain questions in mind. The same is true for experiments since they tend to follow the usual hypothesis-deductive approach. Naturally, this limits the degree to which you can find new relations since the total number of patterns hidden in the data will be limited by the questions that are asked, consequently the existing methods mainly allow you to test an hypothesis. While there is some structure imposed on the text by the thread, sentiment mining on the other hand does not have this limitation since the researcher does not impose the domain of analysis. Any restrictions come from the subjects themselves or to a limited degree from the administrator, and even then, there is considerable room to discuss or describe. So unlike quantitative or experimental data, there tends to be much more free verse that encapsulates the diversity of views and feelings in the sample. Searching through this reservoir of information allows us to find relations inductively that we could not even imagine, and in doing so, it complements the standard hypothesis driven or deductive approaches used in economics.

Under the anonymity and freedom of expression afforded over the Internet, people can be more honest than they are in surveys. Indeed, these issues can dramatically impact the integrity of the data, especially for taboo subjects or those that are ruled by consensus or groupthink. A recent example is the case of Dutch and European surveys regarding race and immigration. Earlier surveys showed the Netherlands to be among the most tolerant societies, significantly more so than culturally similar Flanders. However, questionnaires that are more recent suggest that the differences are not there, something that researchers

had long suspected. They believed that the Dutch ethos prevented people from answering sensitive questions truthfully. Under these conditions, the data collected was misleading and the limitations of survey methodology meant that important research questions remained worse than unanswered.

While it may be possible to get truthful information through mining when you cannot get it through any other means, there are several caveats. Unlike conventional statistical data (which tends to be a small, clean, identically and independently distributed sample (IID) fixed in time), this data is dynamic, excessively large, dependent, and may not come from the same distribution. Therefore, even though it may accurately depict the author's views, it may not be representative of the population since the users of large networks, like the Internet, still do not represent society as a whole. Consequently, there is a tendency to over-sample and under-sample certain demographics, and unlike in a random-sample, this will be true even as the sample size becomes very large. In the end, the data dependency could give invalid standard estimates, even non-conventional estimation procedures like neural networks or state-space search algorithms would face the same problem.

Naturally, a host of sample selection problems would be faced when mining happiness, just as they are with surveyed data. Selection bias, which happens when you choose whom to include the sample, could be a consequence of the crawling and classification procedures described above. Since not everyone is a part of the network and because many users do not express themselves, non-response error is an issue that needs to

be dealt with. In classical surveys, little is known about the non-respondents, but there is a great deal of information known about those that answer the questionnaire. Unfortunately, detailed information about Internet posters is even more limited because online content is being continually updated, the posters (i.e., population) may be changing, and because there are inherent measurement errors—the sample will tend to be non-stationary. Even though it is an important consideration when doing dynamic modeling, in standard econometric analysis this is a less pervasive problem as the data tends to be static. One solution to these types of problems is to consider them when constructing the model, as they do in quantitative sociology where they exploit notions of total sample error. In statistics, the analysis is usually limited to just sampling error, which is a limited portion of total sample theory. Another solution could be to keep track of when each record was stored, by doing this it might be possible to trace population changes over time, and then find ways of correcting for it.

Statisticians working in repeated measurement, time series analysis, and survey design have been dealing with these issues for some time, and they have made headway by developing solutions like data weighting. In principle, these techniques could be applied to mined data, but in practice, this may not be easily possible. Since many of the messages are anonymous, it would be difficult to demographic data on the author. To some extent, this could be mitigated; for instance, if some of their personal information is written in the post or is available through other means like the service provider. Psychological research has faced a similar problem. In North America, experiments and surveys

were often conducted on undergraduate students, who tended to be middle to upper-class males of European origin, so results could not be extended to the population. Nevertheless, psychologists produced a large volume of valid research for their sample base, which has been the steppingstone for a broader program. In any event, this is an issue for nonrandom samples only. One way of getting around the problem would be to collect a large number of writings from known sources, as they do in anthropology, in this way you could approach a random sample by selecting the subjects accordingly.

Collecting responses this way would also avoid multiple counting. Online, someone could express the same opinion on more than one site using the same or different pseudonyms, this repetition could be difficult to detect given that there are some very common handles and that it is hard to verify someone's identity. There are some simple heuristics that could reduce this error. On the same thread, site, or network of sites, you could assume that one name refers to one person. On systems that require a unique user ID this would be true, in other cases it may be a reasonable approximation if you assume that in a local area there is unlikely to be duplication and the smaller the area the smaller its probability. Another correction method could be to keep track of all of the ID's and the number of times each one posts a message. Weighing each extracted message by the inverse of the number of posts for that handle would account for excessive repetition by someone, but it could introduce errors if the person is posting a large number of different messages or if there are large number of people with that pseudonym who post very few and different messages.

Of course, making a correction implies that there is a statistical problem. Since our objective is to assess the degree of sentiment, multiple counting could be justified on the assumption that the more vociferously and the more often someone posts the more intense his feelings. So multiple counting may not lead to a bias, instead it would capture an important quality—passion.

With the entire world to draw from, the resulting dataset will have an enormous number of records and likely more variables than any in existence. The sheer file size resulting from this breadth and depth will bring with it a host of computation, analytic, and test related issues. With potentially terabytes of data, the typical economist's computer will not have enough RAM to hold it all for over a decade. This raises issues of how to store and analyze it. Instead of having a single file that is read into the computer's memory, it will have to be read piece by piece as computational resources become available. One way of organizing this would be to break the grand dataset into smaller files on the same computer (or even across many via a network) that point to each other. This way there would be no need for the analyst to keep track of a large collection of files or to load them in the correct sequence, there would be a seamless transition between one segment and the next. However, it would significantly slow down the analytic process and the lengthen the time it takes to do computations; because, accessing the required bits of information will take more input-output (IO) operations that are by nature slower than calls to RAM, and because much more data sifting will be required in order to get the relevant parts.

The internal organization of the data could contribute to the difficulty in accessing and analyzing information, and it could impose a particular econometric model. For instance, data organized sequentially lends itself to time series modeling, whereas a hierarchical data structure would tend to impose a multilevel model. As well, a single level sequential dataset will necessarily be easier to scan than a multilevel one, which will affect computational time and the ease of analysis. In either case, a large dataset that does not fit into the computer's working memory tends to render traditional econometric analysis techniques less effective.

Exploratory analysis would become cumbersome without the ability to see the entire dataset at once. Cleaning the data is particularly important the larger the number of records because even a small percentage of errors will translate into a numerically large number of outliers. Since we are dealing with potentially tens of millions of entries and possibly a sizable percentage of errors given the way the data is collected, outlier detection is critical, especially if you want to uncover weak relations. Sadly, conventional methods like bivariate scatter-plots become useless for an ultra high number of points and variables; since the data points are spread throughout the files you cannot place all the data points on a chart easily, and the sheer density of points can make the graphs unreadable.

The solution is to use new techniques that have been developed in other fields and to try to ensure the quality of the data from the source. Normally it is difficult to check the source since this information is not kept in the file; but this is trivial problem that can be solved easily, although much more

storage space would be required. However, since the file will tend to be too large any ways, that should not make a difference. At any rate, source-tracking information could be kept in a separate file with just markers to indicate which record it belongs to; this would keep the data file size from ballooning. Once the final raw dataset is ready, we could apply adaptive or sequential techniques to find solutions. These may require iterative procedures or other algorithms when it is impossible to generate closed form solutions. In some cases, even this could fail if the program does not halt. Under that scenario, you could try to build a model using a smaller representative sample of the data. Then, once the structure of the model has been determined, you could use the remaining portion of the data to test it or update it.

The point of all this is to automate the process so that a huge volume of data can be cleaned, reduced, and analyzed in a short period of time, and because traditional methods fail for such large files. Indeed, if real time data analysis is a must, then automation is the only way to go. These computer programs look for patterns; some of them detect outliers, and others try to uncover relationships and build models. To some extent, this is already being done by existing statistical applications like SAS; there are modules that help you select which variables to keep, and there are tools that organize and present the data in a meaningful way, as when you identify clusters. These still require a significant amount of human interaction, whereas the ones for happiness mining would automate data analysis even further.

In order to do this, a fine balance needs to be struck between generality and specificity since the sheer volume of

patterns that could be uncovered would make analysis difficult using the usual econometric methods. If the patterns are precisely formulated, then you will tend to get well-defined families of solutions; limiting the possibilities in this way allows the computer to cope well with the huge volume of data and limitless possibilities, but this is at the expense of finding new and interesting relationships. This could be operationalised by coding specific archetypes and things to look for, or by setting cutoff values for various properties like the conditional probability. Another approach to keep over-fit (in this case, the number of potential solutions) to a minimum is to penalize goodness of fit measures by the number of possible solutions in the class. This would reduce the number of selected patterns, but it may not sieve out the best ones since their characteristics are not being selected. The ultimate solution may be to abandon the probabilistic framework altogether and use a series of rules to select the best candidate solutions, this would select patterns based on some underlying logic rather than expedient calculation. Regardless of the method, automated tools will tend to find fewer but higher quality possibilities when tougher pattern selection criteria are used. When very general selection parameters are used, a large number of possible patterns may be isolated, but they may not be very insightful (or worse, spurious) and the process may get bogged down with information overload. The advantage would be that you could stumble upon unthought-of relationships; which could yield better models by adding to our understanding of the phenomenon's underlying structure and causal relations, whereas the more specific programs would tend to be limited to the roles of data description and forecasting.

These uses bring into question the limits of automation. Clearly, it has an important role in reducing, cutting through, and searching for information in data. However, the procedural rules that allow you to do this do not necessarily lead to a meaningful outcome, and the technical features of the model do not necessarily make it a good one. The distinction between statistical significance and practical significance bears this point. The premise behind hypothesis testing is that by having enough data you can determine, within reasonable type 1 error², whether to reject the null hypothesis (you never accept the alternative because inductively you can never prove something, you can just disprove it). Ordinarily, the trick to doing good econometric tests is to get enough good data to reject the null. The problem is that many of the tests are sensitive to sample size—they become too sensitive to small effects when the sample becomes large enough, often greater than 200 data points. This will certainly be the case with data that is collected over the Internet, so the usual trick will be wholly insufficient, because, the underlying principle is not that there is an effect—it is that the measured effect is much more than the noise in the data, which may be substantial when mining happiness. Thus, in this particular application, there are two effects that call into question the validity of hypothesis test: sample size and noise.

Regardless of what the true sample size is, a common way of controlling sensitivity is to use a fixed value (often set to 200) for the sample size when

² This is the probability of incorrectly rejecting the null hypothesis when it is in fact correct. Usually the 1% level is considered marginal for the natural sciences, 5% is the gold standard in economics, and 10% is acceptable in general for the social sciences

calculating the p-values. This keeps them from overshooting the critical values and provides a similar basis for comparison, so that samples of different sizes can be compared on their merit without being unjustly influenced by the differences in their sample sizes. The hypothesis test may indicate that there is a major difference between the null-change and alternative hypotheses, which suggests that the null should be rejected in favor of the alternative. While this may be true on econometric grounds, one should be wary of rejecting the null hypothesis based solely on this, especially if it has been well established. Instead, you must check to see if the result is economically significant, that is, does it conform to existing theory? Does it take the theory into a new direction? And, is the result meaningfully different from the null hypothesis? Computer programs in general tend to be very good at discovering technical features of a pattern, calculating, and conducting tests, but they are not the best at evaluating the results. Instead, we need to go beyond the mechanical analysis and mathematical structure of the model by showing the output to an expert, so that he can determine if the results are substantive or superficial. One thing to bear in mind is that researchers have a tendency to put the results in the best possible light, so expert opinion can still be biased. To minimize this effect, the results should be evaluated by an independent expert, and a set of clear criteria should be developed ahead of time so that the results can be evaluated dispassionately, indeed some tests in ANOVA mandate that.

An overview of sentiment extraction

As the flowchart in figure one shows, there are a series of steps on the path to extraction, which involve running automated programs, doing manual analysis, searching, cleaning, and finally getting at the critical information. Since the information is spread throughout the Internet, every element of this global network acts as an input to the crawler, which is just a program that explores every node (e.g., web page) on the network and sifts out relevant data. This is then further refined (e.g. by cleaning up the hypertext, tagging key words, isolating negations, and clarifying abbreviations.) in order to get it ready for processing. After preparation, the raw data is scanned in order to see if there are any indications of sentiment. If there are, those parts are isolated and sent to a classifier and quantification module that not only determines if the sentiments are useful as happiness indicators, but it also assigns an appropriate weight to them. The refined data is then organized in some meaningful way (for example by country), analyzed by both man and machine, and the results reported.

Classification

There are a number of methods for sentiment classification but they can be broadly divided into statistical, informatic, and linguistic approaches. In the first classification, the focus is on using metrics to select relevant terms or units, the second is based on computational methods (like machine learning), and the third relies on the construction of the text, vocabulary, collocation, or prepared dictionary lists to select emotive passages. While it is possible to use one or the other exclusively, the best results occur when one filter feeds into the other

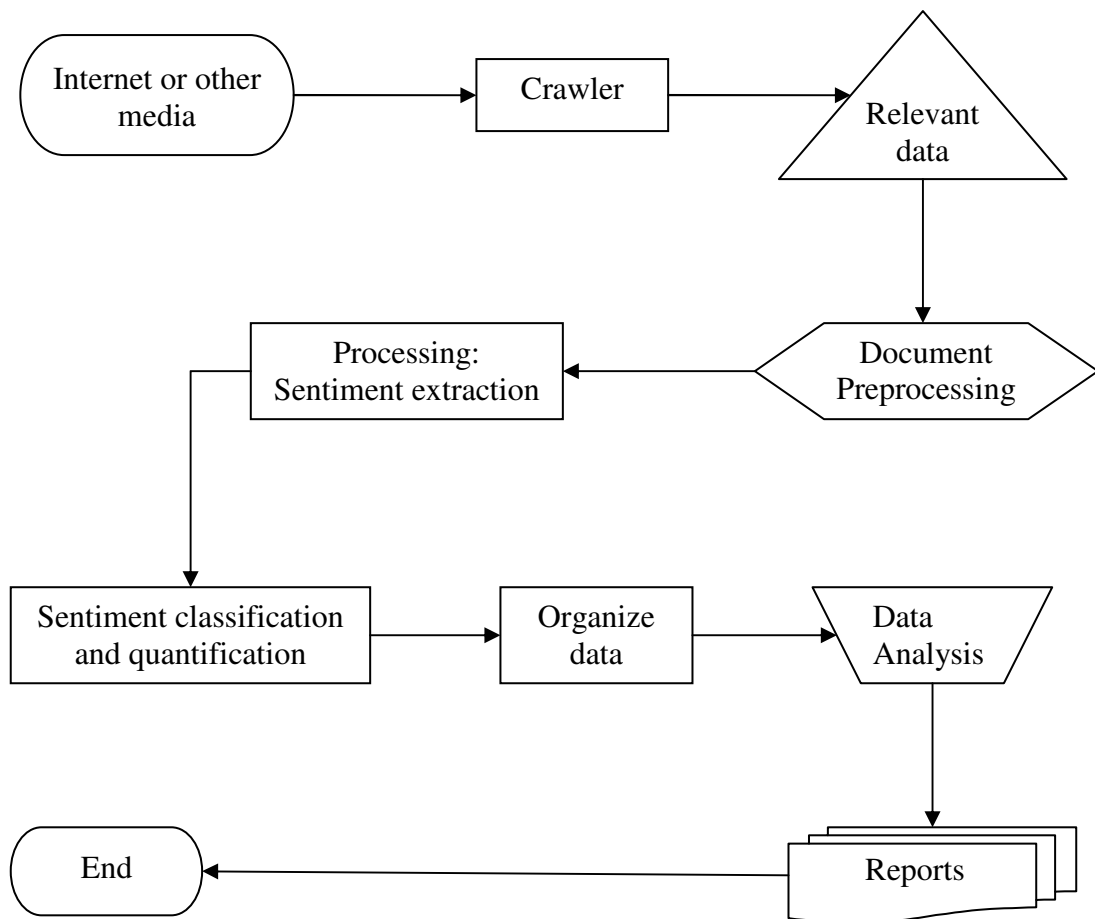


Figure 1: The sentiment extraction process

since all have their own weaknesses—statistical methods require large amounts of data and still produce course results, informatics requires computational power and may not perform better than humans do, and linguistic methods tend to introduce subjective bias. Recent approaches start with sieving the documents through a linguistic classifier since the reliability of statistical measures increases when you use them on a set of more likely candidates.

Graphical methods seem the obvious place to start when looking at classification methods, after all, they help the human mind visualize and intimate detailed information; as McLuhan would say, they are an extension of the eye. A simple time chart can display the number of

positive or negative sentiments overtime, but the real power of graphical methods comes to life when multivariate data is represented on a lower dimensional plane. All but the simplest and most ordered data presented in other forms, like tables, can be difficult assimilate; whereas these projections encapsulate a lot of information and present it in a way that people can quickly and easily consume. The power of visual methods is enhanced further when the graphical software lets you interact with the data, for instance by picking up which points you are interested in or letting you change the orientation. Nevertheless, overtime staring at images can cause your mind to shut down, in which case you could miss interesting artifacts in the data. Luckily, this can be compensated for

with computers, programs can be written to find direct your attention to interesting spots by either preprogramming your criteria or by using machine-learning techniques to learn new patterns dynamically.

One should bear in mind though that there are two things being classified, one is the text as a whole; this is the ultimate goal. The other is an intermediary step, the selection of individual sentiment laden terms or units. These are used to build a lexicon / reference list or otherwise select the relevant texts for analysis. In a lexicon, each word or phrase is added to a specially created dictionary that must be built for each application. While this can be used as a first step in the classification process, it is not the best way to calculate precision (the percentage of extracted terms that are also on the reference list) since the system can extract the correct sentiment word even though it is not on the list, in which case it would be rejected. Nevertheless, this is the most basic component of the classification system and many metrics rely on it. There are several ways of constructing the reference list, some of which are better than others are; these are discussed below.

Experts can be used to classify a text directly or to validate the results of an automated system. The advantage of using human analysis is that it makes it much easier to analyze complex texts that involve several subjects, features, or sentiments. However, using a human expert to go through all the data available would take too long, and in the end may not give much better results for simple texts. Ones that are more complicated, though, still need someone (and ideally many people), to either validate the algorithm's results or go through the semantics entirely.

Since terms tend to be somewhat vague, each expert needs to be trained on their meaning; in any event, they tend to make their evaluation based on their own intuition, external factors, and their experience with terms that they have already evaluated.

This in itself is not necessarily a bad thing, but the tendency, under traditional truth-conditional theories (these treat all members of the category as equal so there is no notion of partial membership) is to try to eliminate the variation between experts since this is considered an error. While the evaluation differences can be reduced for a particular team, independent teams continue to have their differences, which suggests that there is a structural difference in meaning—so by trying to homogenize the interpretations you are actually destroying information and introducing errors. Fuzzy classification is based on the idea that the boundaries of a set are not sharp. Instead membership transitions from one category to another gradually, which implies that some words are more useful than others are, and that disagreements in the interpretation of words among different people can be used as a way to filter out terms that are not central to the category. Andreevskaia and Bergler show that 21.3% of the words found in different lists can obscure the classification process, with the accuracy ranging from 52.6 percent when using lists with more ambiguous words, to a high of 87.5 percent when more definite sentiment laden words are used. Moreover, they find that the boundary between sentiment and neutral words accounts for 93 percent of the misclassifications. This implies that, unlike with non-fuzzy category boundaries, large and random lexicons may not be representative of the population of categories.

Nevertheless, through expert validation it is possible to calculate the algorithms precision by comparing its extractions to those that have been kept by the experts. However, there are some limitations when relying on experts. For instance, it is impossible to evaluate the system's recall ability (the percentage of terms in the lexicon that have been selected by the algorithm.) when experts are used alone because there is no reference list for comparison. As well, since expert based validation methods apply directly to the results extracted by a system they are inherently non-scalable, after all the results of a new algorithm classifying the same text would have to be evaluated again, as would the same algorithm on a new text.

Some automation would help overcome these limitations. Perhaps the simplest relies on frequency / word count, which filters out messages or selects sentiment words by counting the number of relevant units. For example, you might look for messages with emotional words and for words that describe intensity (like depressed, ecstatic, sad, content, crestfallen, satisfied, very, somewhat, little, great, amazing, moderate, etc.), and select messages that have at least one of them (or some higher cutoff, which would give more discriminatory power to). This can be easily implemented by developing a specialized lexicon and comparing it to the message's content. However, it may not accurately capture the meaning of the message, especially as vagueness (in other words, the less the message is about something) of the message increases.

Fortunately, related concepts, the gross and the net overlap scores, may overcome this to some extent. These

count the number of times a word has been classified as either positive or negative by an algorithm or an expert. A lexicon created based on these criteria would tend to contain the highest frequency classifications, and hence would tend to contain terms that have more sentiment content so that issues of vagueness would diminish. While the gross overlap measure does provide more information than simple frequency, it can still be somewhat prone to the problems of vagueness when frequency is low and when the word is classified with both polarities. Fortunately the net overlap measure (the number of times the term is classified as positive minus the number of times it is classified as negative) (NOM) adjusts for this, even though at low frequency it may not be effective.

Andreevskaia and Bergler show that less ambiguity (i.e., greater net overlap score) corresponds with higher classification accuracy. They find that NOM scores of zero (i.e. for categories composed of neutral sentiments or those that mix positive and negative scores equally) give an accuracy of 20%, those with a zero score but only containing positive or negative terms (i.e., they are less ambiguous) have 57 percent accuracy. In general, accuracy increases with higher absolute scores, exceeding 90 percent for NOM scores above seven. In order to use the gross and the net overlap methods effectively over different corpora a list of high frequency sentiment words would need to be generated; naturally, this would be a more computationally expensive and time consuming process than using simple frequencies, especially if experts are involved. Even though simple word counts have some limitations, Pazienza et. al. find that they make the best compromise for creating lexicon's for sentiment extraction, which is supported by Evert

& Krenn since true terms tend to have high frequency and because the metric is computationally very light. Not surprisingly, Krenn finds that statistical measures do worse when applied to low frequency terms.

Linguistic methods can be very effective as the first level filtering system. Several techniques fall under this umbrella including information extraction, semantic orientation, and natural language processing. The first method takes text from disparate sources and stores it in a homogenous structure, like a table. The idea is just to analyze the relevant portions of the original text, not necessarily the whole document. The second can be applied to individual words or the entire passage. A word's orientation depends on how it relates to other words in the text and on its relation to strength indicators; these include words like great or horrible. By adding up the semantic orientation of each word, you can calculate the orientation of the entire text. The caveat is that this method is based on the assumption that the text pertains to only one topic. Consequently, passages containing many different subjects would give an incorrect semantic orientation score, and it could lead to them being assigned the wrong polarity.

Natural language processing improves on the situation. It usually implies parsing (or at least part of speech tagging), and mapping the results onto some kind of meaning representation. One benefit is that it allows sub-topic sentiment classification; therefore, you can pick out particular things that make someone happy or unhappy. Another is that, by using the information the grammatical structure to determine what the referent is, it can reduce the errors made when associating the sentiment to the topic.

This can be done by: isolating phrases that contain definite articles, using prepositional phrases and relative clauses, and looking for particular structures of verb/noun-adjective phrases or their particular location in a sentence. By extracting only those portions of the text that can grammatically contain the concept, you automatically exclude words that would otherwise pass through your filter or be mistakenly recognized, so it makes the selection procedure execute faster (since there's a much smaller input), and it keeps errors lower.

Kgeura and Umino point out that text content can be grouped into two domains, unithood represents how strongly collocations (the elements of the unit) are related to each other, and termhood expresses how tightly coupled the term is to the underlying concept. Although this classification scheme allows you to search for good sentiment candidates based on two linguistic dimensions instead of just one, Pazienza et. al. find that measures for both of these seem to select similar terms, especially those that are ranked highest. Since they contain essentially the same information, it may be possible to generate good classification lists using just one of the dimensions, which would reduce the computational overhead. However, which one is better is unclear. In Pazienza et. al., 2/3's of the effective measures, including frequency which is the most effective one, are geared towards termhood, but, they find that the log likelihood ratio does not effectively recognize sentiment in unithood, whereas Dunning and others do. So more work needs to be done to see how effective the likelihood ratio is. If it turns out to be less effective, then algorithms could focus more on termhood. If it turns out to be an effective predictor, than perhaps both

termhood and unithood should be retained.

Quantitative indicators of the unit or term's relevance can be broadly classified as heuristic or association measures. The first tend to lack a strong theoretical basis, instead they are based on empirical observations or logical intuition. Consequently, they cannot guarantee a good solution, but like frequency and C-value, they tend to give reasonable answers. In fact, they can give better solutions than their theoretically sound counterparts, one explanation is that they correct for an underlying weakness in the theoretical measure. Take the Mutual Information metric, since it tends to select words that occur infrequently in texts, its power to discriminate meaningful sentiment words is reduced. However, by cubing the measure there is a tendency to overcome this weakness, at least in the highest-ranking terms.

Association measures tend to quantify the degree of collocation between words, which is based on the observed frequency values in the unithood contingency table. Unfortunately, from a practical standpoint, estimating these frequencies is not a trivial matter. As the number of variables increases, the overall probably distribution becomes more complex and, for categorical variables, it becomes impractical to estimate the individual cell counts. However, the Bayesian network offers a solution. If you can summarize the distribution in terms of conditional probabilities, and if a small number of these are positive, then the number of parameters that need to be estimated declines immensely and the precision increases.

Estimation issues aside, in order to extend these inferences to the population as a whole, a random

sample model should be used to estimate the collocation values (i.e. the elements of the population's contingency table). Dunning points out that if the collocations are mutually independent and stationary (i.e., the probability of any particular word coming up in the text is constant), then the elements for a 2 unit collocation can be taken from a Bernoulli distribution with each element of the table having its own probability of occurrence, each of which need to be estimated. Degree of association metrics directly estimate these parameters using maximum likelihood estimation. Naturally, this method is replete with estimation errors, especially when frequencies are low. To reduce this effect one can rely on another class of metrics, significance of association measures, which use the null hypothesis of independence. In other words, they assume that there is no relation among the words in the collocation; so, the cell probabilities are the product of each word's independent marginal probability, which are computed using maximum likelihood estimation. Under these two conditions, expected frequency of the cell is equal to the mean of the binomial distribution.

Empirically, measures based on frequency and significance of association outperform degree of association metrics in terms of precision. The difference between these measures is greatest at the first recall percentile where the highest measures range from precision values of 0.6 to 0.75 and the degree of association measures range from 0.3 to .45. However, by the tenth percentile, they all converge to the precision of .5, with the frequency and significance metrics consistently giving higher precision values. This suggests that the discriminatory power

lies with just the highest-ranking words, which tend to have the highest frequency scores too. This is similar to the notion of centrality, which represents how close an element is to the core concept. Centrality can be measured in a number of ways. Two possibilities are by assessing the number of semantic links that a word has with other words in the category, or by measuring the amount of disagreement there is on the words membership (for example by counting the net number of times a word is classified as having positive or negative sentiment), which is similar to this case

Words that carry the most relevant emotion would necessarily have high centrality, therefore they would tend to be used more often, so the extremely low precision values of the degree of association measures can be explained by the fact that they rely too much on low frequency terms. This is borne-out by the fact that the frequency metric (and those dependent on frequency) tends to consistently have the highest precision, mutual information (which tends to select low frequency terms) has the lowest, and its cube (which selects more high frequency terms) does much better than it at lower percentiles.

This suggests that, among these measures, frequency laden measures will tend to extract sentiment most precisely, and that methods that rely on the null hypothesis of independence (like the t-score or the log likelihood ratio) should be preferred to those that only try to approximate the probability parameters (like the dice factor). Computationally, the implication is that, primarily, it may be best to use the frequency metric as the basis for term selection—it appears to be the most effective and needs the least

computing power. Perhaps though, the other measures could be used to validate the frequency-based list in an attempt to improve precision further by isolating the remaining outliers, but this idea needs to be investigated further.

Happiness, social contagion, and influence

One can imagine that actions are taken in response to a broad set of influences. The most basic are preprogrammed responses. These occur automatically in response to a physical need or a change in the environment. Primitive drives, like hunger, propel humans to search for and consume food, and in dangerous situations people are subject to the fight or flight response. Another influence, can be broadly classified as social contagion. Gatherer says that in this process, behavior spreads like an infection when agents interact. Think of the case where people are at an event, as soon as one person leaves many others do too. A third device is social control, where agents are influenced to follow norms or commit other actions. Even though the last two mechanisms are a form of social propagation, none of these three instruments fall under the umbrella of social learning (which includes social facilitation and imitation) because no learning takes place. Instead, the response is relatively automated, and it does not cause the production of a mental model nor is it directed by one.

Social propagation refers to a group of mechanisms that lead to the transmission of behavior. As shown in figure 2, it can be broadly separated into cognitive and non-cognitive processes that can be arranged in a sequence, with the simplest and least cognitively challenging ones on the left side and the most complex ones on the

right side. Social contagion and control both fall into the non-cognitive portion since thinking is not necessary, while social facilitation and imitation are considered cognitive in nature. Notice that social learning does not include the whole of facilitation and imitation. This is because the two encompass a wide variety of behaviour transition mechanisms, some of which may not require cognition.[Insert Fig. 2]

Social contagion occupies the left most portion of the behavioral propagation space in figure 2, meaning that behavior can be transmitted with very little cognitive effort on the part of the receiving agent. Consequently, the behaviours that can be learned like this are usually simple and short lived, meaning that the agent's mind does not create and store a mental representation of the behaviour that can be used again. Usually the transmitted behaviour is purposeful just for that moment, although similar situations in the future may elicit the same response. For instance, it has been shown that, like emotions, stretching and sneezing are contagious, one person doing this in a room will start a chain reaction. Coined social release³, these types of actions happen spontaneously and usually the agent is already familiar with them, but they need to be started off through a social interaction according to Heimann. Strictly speaking, according to this perspective the contagion cannot be learned because the agent already has the ability (e.g. to yawn). As well, since the behaviour is temporary, it is assumed that no mental representation is created, so the knowledge base cannot have been updated and

³ The work release is used because it is assumed that the agent already possesses the ability, but because it isn't manifested until after the interaction, the interaction was responsible for releasing it.

therefore future behavior cannot change. This perspective follows a strictly behavioral philosophy *a la* Skinner, under whose guise critics would argue that a deeper understanding of the process is unwarranted because there are no long-term behavior consequences.

Over the last decade, social interactions have moved into the virtual realm, with people interacting through email, chat rooms, blogs, video, etc. Happiness and other emotions could be transmitted through these media, but perhaps not as effectively as in the examples given above because fewer senses are involved and the means of expression are limited. This smaller effect size would be more difficult to separate from the noise so it would be more difficult to measure. Even so, the psychological phenomena mentioned above should be unaffected and we should be able to see, measure, and model behavior propagation and influence. It has been shown that feelings can be transmitted from person to person, often in the most unassuming way; if you smile, others around you do too, laughter makes things seem funny, and ill-temper travels from person to person like the common cold. If similar effects can be transmitted online, then we should be able to see evidence of this, and we should be able to measure if some sources are more influential than others are.

The notion of influence suggests that the beliefs and actions of people can be guided by external forces, and that some hold more power than others do. Two components of influence are credibility, which refers to how much people trust the source (both the content and the person), and authority, which describes someone or something

with power or expertise. The CEO of company has authority over his personnel by virtue of his position, clerics hold sway over their parishioners through moral suasion, and professors can influence because people trust their knowledge. Regardless of the reason, credibility and authority have the power to influence the things people do, say, and think, but what about feel? Research shows that religious people are happier than atheists are; this gap could be explained to the extent that religiosity is the product of environment and authoritative persuasion. If this is the case, then visitors and contributors to authoritative and credible online sites could have their emotional state altered by the content.

The power of influence is somewhat constrained since it tends to be topical, temporal, and polar. Some of its constituents, like authority and credibility, only hold sway over a particular or cognate subject matter. After all, a renowned expert in biochemistry may exert great power of mind over biologists and chemists but hold little clout with anthropologists. The very nature of influence suggests that it moves people from one point to another that is significantly different. For instance, if you already hold a very positive sentiment initially, then you do not need to be influenced much more; the degree of the sentiment may change a little but its direction will be the same. The real impact would be to turn your feelings around, as could happen, when you move to a new city and get new circle of friends with the opposite opinion. The third constraint is temporal; the power to move people is fleeting, so a truly influential online source would be long lived.

Using the techniques described in this paper, the sentiment of the content can be analyzed, and a contributor's change can be measured since he can be tracked to some extent. It is the lurkers who can pose a problem because in many cases little or nothing is known about them or just how many there are. This is usually the case for content, such as web sites and blogs that can be accessed actively without logging on. In other cases, the data is not so dark since some of the lurkers' qualities can be measured. For instance, the number of people subscribing to a mailing list is known, and so are the number of contributors and their handles. The difference between these values would give a simple estimate of the number of lurkers, and the registration records may give background information on the subscribers. By keeping track of things like: when the person subscribed, who is posting and when he contributed, the sentiment of the combined logs, and the sentiment of each post, you could in principle estimate things like how long a person has been lurking, the sentiment he has been exposed to during the period, and the sentiment he finally revealed. If the revealed sentiment of lurkers is the same or stronger than the overall sentiment of the articles it could imply that he was influenced, or that he continued to lurk because his views were similar. In the first case, you expect an upward trend in sentiment with respect to the time spent lurking.

The important thing is that there is an enormous amount of dark data that is difficult to collect or estimate. Not all people write messages, but they still have opinions and feelings so we need to develop methods to account for this. Page rank tells you how popular a particular piece of information is, the higher the rank the more people have viewed it. This could be used as a

measure of influence, and it could be a measure for the number of people who hold a similar opinion. Gathering statistics for networks of sites would also give similar measures. Counting links, looking at the amount of traffic / readership, doing temporal analysis, and putting the network data into a graphical form would help you account for some of the dark data. This could be used as a way to reduce the error in your estimations, for instance by using page rank as a weight in the sentiment score. Joshi constructs a model of influence that infers the degree of influence and sentiment of political blogs. He finds that polarity after trust propagation, in the worst reported case rises is from +1 to +1.370 and in the best from +1 to +9.570, which suggests that emotions like happiness may be influenced by online content and that we may be able to measure it.

Conclusion

The decline of mainstream media like television, and the rise of new media like social forums points to a new direction of information exchange, one that is more conducive towards measurement than its not interactive predecessors are. Yet to date, the wealth of information has remained largely hidden from scientific inquiry, even though it is abundant, inexpensive, and current. Some work has been done in this field, often by computer scientists in their effort to develop better technologies to exploit the dark data. These mining techniques have also been applied in a number of disciplines including finance, ethnography, and political science. However, comparatively little has been done in the field happiness.

While there are some challenges in terms of insuring data quality and developing sound measurement techniques, these are surmountable.

Some of the most important technical, linguistic, econometric, and informatic hurdles have been discussed above and ways of coping with them highlighted. Given the challenges involved in creating and then working with such a large and diverse dataset, a hybrid approach that joins several classification methods (like statistical, algorithm, linguistic, and human opinion) appears to be the best approach. In principle, there seems to be little reason why this line of inquiry should not succeed, especially given the promising theoretical and empirical results that have been highlighted earlier.

Last section discusses how happiness mining could be used to study how influence and social contagion influence individual happiness. The fundamentals of these concepts are explained and some approaches for data collection, metric creation, and analysis are discussed. Results from political blogs suggest that influence via trust can drastically affect the sentiment index, suggesting that perhaps a similar effect could be seen in online revealed happiness data.

In the end, mining happiness gives economists a rare opportunity. Ordinarily, they are forced to rely on existing data sets over which they have little control. Even when they can design the survey, the questions are guided by existing hypotheses, which limits degree to which new avenues can be explored. With text mining, the tables are turned and the designers of questions can now become searchers for answers, letting the data reveal the model, rather than implicitly imposing one on the data.

A common misnomer is that data mining is somehow unscientific and that it does not reveal facts of

substance. In reality, there is a strong scientific method underlying the mining procedures, and these are being supported by advancing technology. Good mining starts with a clearly identified problem, a well-defined objective specification, and a set of objective and subjective criteria that can judge the merits of the find. This forms the foundation of a thorough search, which balances general and rigid pattern recognition in an attempt to filter out random patterns while isolating those that reveal the underlying structure. Over time, this is bound to yield some novel results that could drive theory in new directions. As such, it complements the usual hypothesis-deductive-empirical approach in economics and other sciences, and surely no one can disagree that having one more tool at your disposal is an advantage. Certainly, if done correctly it has the real potential of validating existing research or suggesting new avenues of exploration.

Given that there has been limited use of this technique in happiness studies, there are many avenues for further research. One could ask how the connectedness of sites or their structure (e.g., the similarity of the postings) affects happiness. In other words, if the people in your network are happy are you more likely to be happy? Technical issues like the quantification of the noise in mined happiness data are also interesting. More work on developing algorithms that do searching, classification, cleaning, reducing, and analysis is needed. Finally, it could be interesting to see which term classification methods would improve a frequency-based lexicon. Of course, any question that has been asked on conventional data could be asked on mined data, the challenges to see those questions through.

References

- Admati, A. R., Pfleiderer, P.
Noisytalk.com: Broadcasting opinions in a noisy environment, WP1970R, Stanford University, 2000.
- Ahmad, K. and Yousif, A. Visualizing Sentiment in Financial Texts. Proceedings of the Ninth International Conference on Information Visualisation (IV'05). 2005.
- Andreevskaia, A. and Bergler, S. Mining Wordnet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. Proceedings of 5th International Conference on Language Resources and Evaluation (LREC-2006). 2006.
- Antweiler, W. and Frank, M. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *Journal of Finance*, American Finance Association. vol. 59(3), 2004.
- Balbi, S. and Misuraca, M. Visualization Techniques for Non Symmetrical Relations. *Knowledge mining* Vol. 185. Springer, Berlin / Heidelberg. 2005.
- Camillo F., Tosi, M., and Traldi, T. "Semimetric Approach, Qualitative Research and Text Mining Techniques for Modelling the Material Culture of Happiness". *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference*. (ED.) S. Sirmakessis. 2005.
- Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. Identifying sources of opinions with conditional random fields and extraction patterns. *Proceedings of Human Language Technology & Empirical Methods in Natural Language Processing Conferences (HLT/EMNLP)*, Vancouver, 2005.
- Das SR, Chen MY. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*. Vol. 53, No. 9, September 2007.
- Das, S. and Sisk, J. Financial communities. *Journal portfolio management*. v31(4), Summer 2005.
- Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1). 1994
- Eikvil, L. Information extraction from world wide web - a survey. Report No. 945, July, 1999. ISBN 82-539-0429-0.

- Ellison, G, Fudenberg, D. Word-of-mouth communication and social learning. *Quarterly journal of economics*. 110(1), 1995.
- Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France. 2001
- Fyhrlund, A. Fridlund, B., and Sundgren. Using Text Mining in Official Statistics. *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference VIII*. Editor Sirmakessis, S., 2005.
- Gatherer, D. The Spread of Irrational Behaviours by Contagion: An Agent Micro-Simulation, *Journal of Memetics - Evolutionary Models of Information Transmission*, 6. 2002.
- Généreux, M. and Santini, M. Exploring the Use of Linguistic Features in Sentiment Analysis. *Corpus Linguistics 2007*. Birmingham, UK. July 2007.
- Gray, Peter O., *Psychology*, Fourth Edition, Worth Publishers; 4th edition, 2001.
- Heimann, M. When is imitation imitation and who has the right to imitate?, *Behavioral and Brain Sciences*, 21, 1998.
- Izquierdo, J., and Larreina, S., *Collective SME Approach to Technology Watch and Competitive Intelligence: The Role of Intermediate Centers*. *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference VIII*. Editor Sirmakessis, S., 2005.
- Java A. A Framework for Modeling Influence, Opinions and Structure in Social Media. In *22nd Conference on Artificial Intelligence*. 2007.
- Kageura K., Umino B.: Methods of automatic term recognition. *Terminology*, 3(2). 1996
- Kale A, Karandikar A, Kolari, P, Java A, Joshi A., and Finin T. Modeling Trust and Influence in the Blogosphere Using Link Polarity. In *Proceedings of the International Conference on Weblogs and Social Media*. 2007.
- Krenn B.: Empirical Implications on Lexical Association Measures. *Proceedings of The Ninth EURALEX International Congress*. Stuttgart, Germany. 2000
- Muslea, I. Extraction patterns for information extraction tasks: A survey. In: *The AAAI Workshop on Machine Learning for Information Extraction*. Menlo Park, CA, USA, 1999.
- Neri, F. and Raffaelli, R. Text Mining Applied to Multilingual Corpora. *Knowledge mining. Knowledge mining Vol. 185*. Springer, Berlin / Heidelberg. 2005.
- Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2002
- Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M., *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference VIII*. Editor Sirmakessis, S., 2005.
- Šoltés D. New Challenges and Roles of Metadata in Text/Data Mining in Statistics. *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference VIII*. Editor Sirmakessis, S., 2005.
- Talwar, A. Jurca R. Faltings B. Understanding user behavior in online feedback reporting, In: *Proceedings of the 8th ACM conference on Electronic commerce*. San Diego, California, USA, 2007.
- Verma, R., *Creating an Emotive-intensive Lexicon for Social Research using a Hybrid Procedure*. Submitted.
- Wysocki, P. "Cheap Talk on the Web: The determinants of Stock postings on message Boards". WP 98025. University of Michigan Business School. 1998.
- Zaima JK, Harjoto MA. Conflict in Whispers and Analyst Forecasts: Which One Should Be Your Guide? *Financial Decisions*, Article 6, Fall 2005.