RESEARCH ARTICLE OPEN ACCESS

Using Change Point Analysis to Determine Perception Accuracyinsocial Media Opinions

Philip Sallis*, William Claster**

*(School of Computer and Mathematical Sciences, Auckland University of Technology,New Zealand)

**(School of Asia Pacific Management ,Ritsumeikan Asia Pacific University ,Beppu, Japan)

--***********************-------------------------------------

Abstract:

Social media such as Facebook and Twitter have become repositories for extremely large volumes of data representing for the most part, personal opinions and statements of individual users of these instruments for interpersonal communication. They provide a forum for discourse and debate. The strings of words broadcast by individuals are predominantly casual expressions reflecting personal attitudes or sentiments. Due to the pervasive nature of social media and the opinions they proliferate, it is assumed likely that these casual expressions are being relied on by some readers of the discourse, for their decision making influence. This paper describes how Change Point Analysis (CPA) can be used to bring some precision to the plethora of casually expressed sentiment sourced from social media databases. It is contended that where the conjunction of opinions expressed are related to a factual occurrence in time and space, a change point can be established, which indicates the potential reliability on a given sentiment as being a reliable influence factor for opinion formulation or other decision. When related to tourist destination data, the case study for Japan described here is convincing in its analytical result exposition.

---***********************---

I. INTRODUCTION

From time-to-time there is criticism that imprecise or inadequately fact-based opinions are presented in the public press as accurate representations of reality. This phenonenumis no less obvious in social media, which in recent times has become cited as indicative of both commonly held and extreme views expressed as fact. While there may be some degree of reliability in utilizing opinions expressed in social media, the reliance being placed on these sources for fact leads us to seek ways of resolving the ambiguities of language inherent in casual expression and the instances of poorly constructed phraseology or simply insufficient words used to clearly put a personal view when using this informal communication form.

The tourism industry generally is one where experiential data gathered from travelers can influence substantial changes in services, locations and provider behavior. Previous studies by the authors (see [1] for example) has shown trends in traveler destinations and service use, based on such influences as political unrest, climate variation or economic disruption. When conveyed from social media sources these casually expressed opinions, while possibly accurate are usually founded on single or small sample personal experiences. It would seem that the rapid uptake in social media use across all sectors of society worldwide has encouraged a sense of reliability on the opinions reflected in it and therefore, has become a credible influence on decision making for individuals.

Social media and sentiment analysis have been used to model financial markets [2] and political elections [3]. There has, however, been no work in the area of social media analysis. The work we describe here is original in its hybrid use of change point analysis and intervention variables. A similar methodology was used to model electricity prices [4] and Japanese foreign exchange rates [5].

In the research described here, we employed data mining algorithms (particularly neural networks, decision trees, Bayesian methods) to test the assertion that algorithmic analysis of large amounts of social media can provide previously unused data to deliver relevant predictions, just as in the aforementioned research. It should be stressed that the analysis of social media has been shown to have predictive value in fields other than hospitality – fields such as finance (stock markets) [2] and politics (elections) [3].

To illustrate the phenomenon proposed here, this paper describes a case study based on data from Japan that illustrates the nature of social media opinions and their impact on the tourism sector. It outlines an analytical method, change point analysis, and applies it formally to the social media data to demonstrate how precision can be brought to the casually expressed opinions by extracting points in time when the weight of evidence for a particular opinion is supportable across a sample data set such that it becomes a credible decision-making factor.

Japan, with a total non-Japanese resident population of less than 1%, nonetheless attracts a large number of visitors. Seeking destination information, opinions on hotels and sightseeing options has led to a substantial source for social media originated data.

Perceptions of Japan from outside its borders can be monitored through press, magazine, and other traditional media. Blogs and other social media however, are important in reflecting both current perceptions and determining them as the antecedents and the consequents in a causative relation. In this paper we focus on the three major cities to ascertain the potential influence of social media in an attempt to understand the functions of these particular media and their particular focus for the tourism industry.

In Japan, the tourism industry is a primary instrument for understanding foreign perceptions and often this serves as an introduction for person-to-person contact with many foreigners. Japan has awidely held and unique image among the nations of the world and is sought after as an intriguing tourist destination. Its economy is third largest in the world.

Some previous work by the authors with others [8]is described here where social media data was collected, analysed for opinion similarity and coincidence of terminology used, then proposed as an inferential model of opinion trend. Following this description, an outline of the data used for this current paper is given as is the method used to process and analyse it. This includes a novel approach to data analysis using change-point-analysis, which is described in other work previously conducted by the authors and others [9].

The results from the research carried out for this paper are described and discussed then some conclusions are drawn from them and the domain of investigation generally. The significance of this work is proposed as contributing to the quest for finding adequacy in formally modeling the contemporary phenomenon that is social media and its implications as an economic and business driver.

Previous studies – data mining :

Information concerning where to go, what to do and what to expect when travelling has shifted in recent years from being print material based to internet sourced. Online reviews are becoming increasingly useful and used in this respect. As a research topic, automatic review mining and summarizing has become a topic of increasing interest, especially when it can be shown to produce accurate and reliable information. Horrigan [9.5] for example, noted that 81% of

Internet users have done online research on a product or place at least once. Another study in 2007 conducted by Comscore [11] revealed that user reviews has a significant influence on customers' purchase, and that reviews generated by fellow customers have a greater influence than those generated by professionals. Vendors can ascertain, both from current customers and potential customers, information that previously may have beyond their reach. Subjective information related to objective characteristics such as a customer's subjective view of product design may be available in blogs and review sites. In addition knowledge pertaining to political and policy issues is communicated across the web and may be utilized to formulate policy that is attractive and rooted in the public interest [12], [13]. During the 2006 elections in the United States, 34% of campaigners used the Internet to gather information and exchanged views about the 2006 elections online [14].

Accessing and measuring the sentiment accumulated in the vast store of blogs, online publications, social network media (such as Facebook) and microblogs such as Twitter can yield tangible and actionable information for Business, Marketing, Social Sciences, and government. Knowledge of consumer opinions, public attitudes, and generally the "wisdom of crowds" can yield highly valuable information. As the World Wide Web has developed, considerable decision making power over the consumption of discretionary products like tourism has been transferred from suppliers to consumers; there is thus a real need to improve market intelligence and market research for private and public tourism organizations to facilitate timely consumer decision making. Here we explore the development of user generated content about the characteristics and value of destinations through analyzing the use of Twitter and seek to answer whether tweets can be mined for industry intelligence.

Twitter posts may be regarded as conversational microblogs. In the previous work these comments relate to [12] we proposed that these microblogs can be used as a source of sentiment expression and for this study we focused on sentiment expressed towards the travel destination of Phuket a travel resort region in Thailand.

Sentiment mining aims at extracting features on which users express their opinions in order to determine the user's sentiment towards the query object. For the previous study described above, we mined over 80 million Twitter microblogs to gain knowledge regarding sentiment on the toward both Bangkok and Phuket during the 7 months from Nov. '09 to May '10 in order to discover whether this social media reflects and may explain sentiment towards the tourist resort Phuket in light of the political unrest that occurred mainly in Bangkok during this period. In this work we explored whether tweets from the social media initiative known as Twitter can be used to identify sentiment about tourism and Thailand amid the unrest in that country during the early part of 2010 and whether analysis of tweets can be used to discern the effect of that unrest on Phuket's tourism environment. It was proposed that this analysis can provide measurable insights through summarization, keyword analysis and clustering. We measured sentiment using a binary choice keyword algorithm and a multi-knowledge based approach was proposed using Self-Organizing Maps along with sentiment polarity in order to model sentiment intensity. We developed a visual model to express a sentiment concept vocabulary and then applied this model to maximums and minimums in the time series sentiment data. The results showed that actionable knowledge can be extracted in real time. The results were convincing as illustrated below:

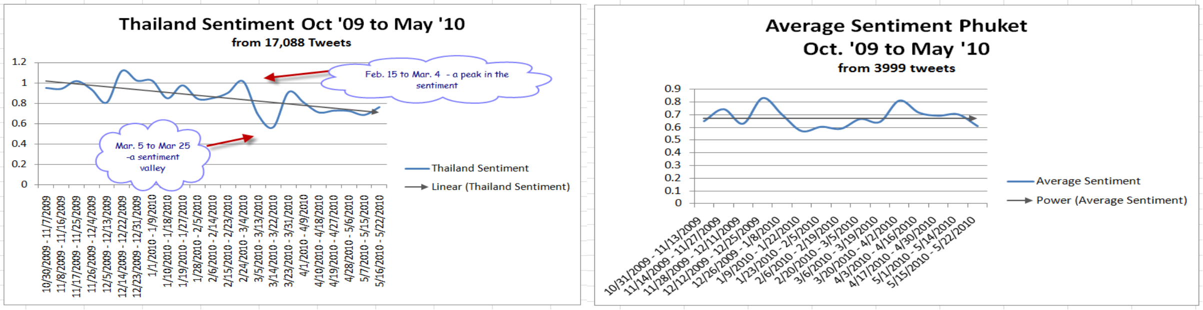


Figure 1. Time series of sentiment tracked through tweets from a data set of 70,570,800 tweets comments or about 20.42 GB data drawn from Oct. 30, 2009 to May 21, 2010

Redshirts	Wars	Violence	Holidays	BBC
followers	govt	army	fly	wounded
troops	guys	violence	holidays	bbc
anti-government	military	district	live	bad
rally	money	china	luxury	explosions
Phuket	security	super	road	TV
redshirts	wars	bookings	special	bank
hotels	work	forces	thb	British
protests		march	villas	fire
		armored	waiting	mafia
		krabi		pictures

Table 1. Time series of sentiment tracked through tweets from a data set of 70,570,800 tweets comments or about 20.42 GB data drawn from Oct. 30, 2009 to May 21, 2010

The Self Organising Map that was produced from the data analysis depicted below, illustrates the relative intensity of the tweets made relating to the period and location being studied:

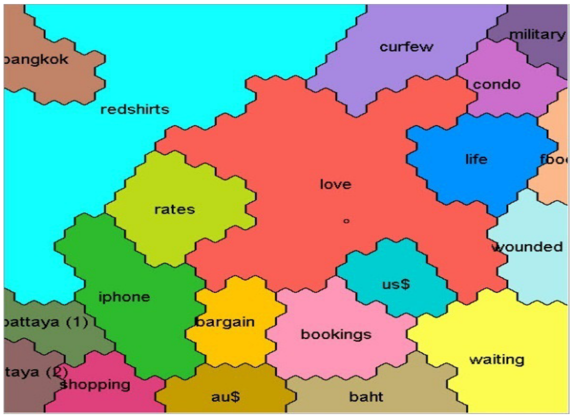


Figure 2. SOM of all Thailand tweets over entire period of data collection: Oct. '09 to May '10.

Previous Studies – Change Point Analysis

Change Point Analysis (CPA) is a formal method used for determining whether a change-of-state has taken place in a set of observed events. There are claims that this method is capable of detecting subtle changes missed by for example, control charts. By providing confidence levels and confidence intervals it claims to better characterise the changes detected. This is a bonus when for instance, tracking events in Nature especially where continuous data is concerned such as is necessary for wind velocity data analysis. This is particularly the case when in order to detect perturbations occurring in the oscillation patterns of airflow because these are potential predictors of wind gusts, which can have devastating physical effects on land-based objects such as buildings and crops. The research referred to for the purposes of this current paper [9 op cit] relates to the prediction of wind gusts using a branch-and-bound algorithm. It considers the need for precision and early detection in wind pattern state-change and examines how fit-for-purpose a change-point analysis method could be for the early detection of velocity oscillation perturbations in a mixed variable analysis of condition change. Wind velocity data is sampled in real-time. The continuous data feed is processed by a change-point analysis algorithm, which has been derived for this purpose. The results obtained can be seen depicted in the figure 3 below, where the red dots illustrate wind gusts and can be related to the other climate factors that are plotted against their occurrence:

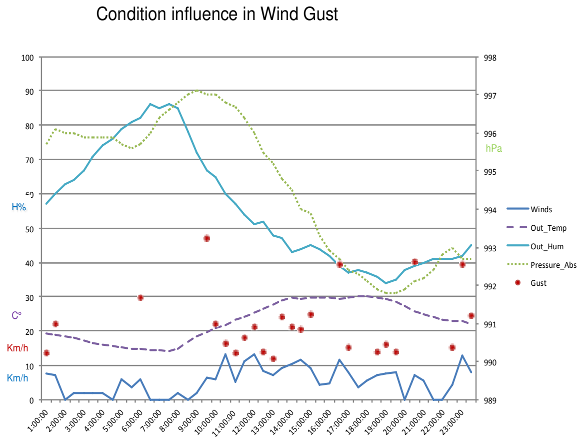


Figure 3 Discrete variables as factors relating to wind gust events

The derivation of the wind gust prediction results shown here were the result of even earlier work by the authors [15]. When overlaid with the CPA analysis results, the cluster of wind gust occurrences can be seen falling within the confidence parameters of the CPA algorithm thus:

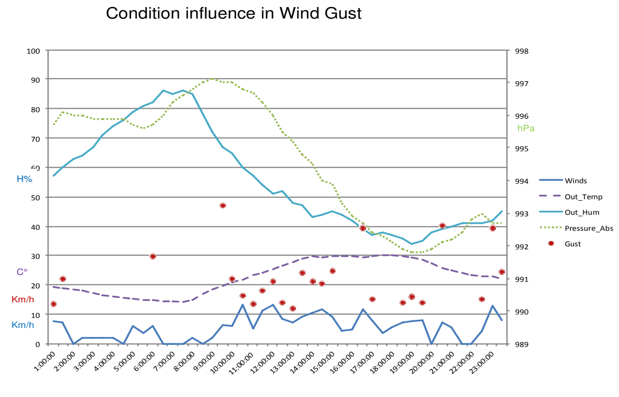


Figure 4 Superimposing the CPA parameters

For this paper, we have taken a blended approach to data analysis first, extracting the tweets and charactering them for their sentiment intensity and then second, identifying algorithmically the confidence parameters that determine when change points can be identified, indicating modifications in intensity values.

The literature relating to Change Point Analysis (CPA) mostly relates to its use as a method in financial modeling and more widely in economic dynamics and trend forecasting. In a recent article describing this application area Taylor [16] claims that although not fully appropriate for use with continuous data through say an on-line real-time feed, the additional confidence level information provided by the change-point analysis (CPA) method when used together in a hybrid approach, produces more reliable state-change models than previously developed. He goes on to say that when analyzing historical data, especially when dealing with large data sets, change-point analysis is in fact, preferable to control charting. He provides a useful illustration of control charting where in Figure 5 below, the red horizontal lines indicate the upper and lower bounds of an individual chart where it is assumed over time no change has occurred. It can be seen here that at the point shown to be in the time period October 1987 that in fact, a change point did occur because it exceeded the upper limit of the normalised parameter.

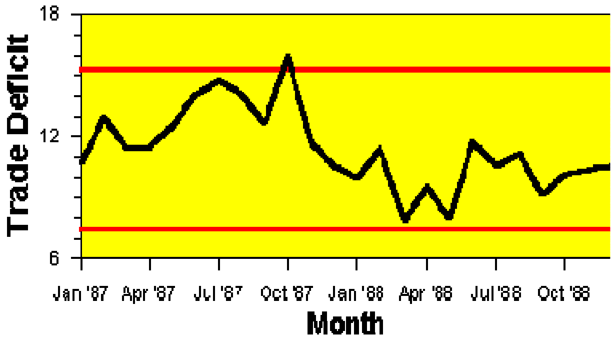


Figure5 Control Chart with assumed fluctuation limits [16 (op cit)]

This is argued to be the case because CPA is a so-called more powerful analyser in that it better characterizes the changes, controls the overall error rate, is robust to outliers, is more flexible and is more simple to use. These observations are attractive to those working in areas other than finance or economics because it may be that CPA is a tool that in one form or another could be used to observe state change considered to be statistically significant. This approach would for example, enhance our understanding of so-called step-change events, more precisely quantified using such non-parametric methods such as McNemar's Test for the Significance of Changes [17]. This long-standing method originating from early (1947) psychometric research is used on nominal data. It is applied to contingency tables of a dichotomous trait with 2 × 2 dimension, which have matched pairs of subjects. It is used to determine whether the row and column marginal frequencies are equal. McNemar [18 (op cit)] calls this derived value marginal homogeneity. It is determined by testing the null hypothesis of marginal homogeneity, which states that the two marginal probabilities for each outcome are the same.

So the marginal probablities are compared thus,

$$pa + pb = pa + pc \text{ and } pc + pd = pb + pd$$

and the null hypothesis isexpressed as pb = pc.The McNemar statistic equation below uses a chi-square derivation in order to produce a binomial distribution of matrix values thus,

$$X^2 = \frac{(|b-c|-0.5)^2}{b+c}$$

If we apply this in an example where the same subjects in a sample are included in a before-and-after measurement matrix (so that they are matched pairs) we can see as in Table 2 below that we have the potential for observing a binomial distribution. The data depicted here relates to patients who were diagnosed with a particular disease and then treated with a prescribed drug. The effects of the drug before-and-after for each patient were observed [16 op cit] and [17].

	After: present	**After:** absent	Row total
Before: present	101	121	222
Before: absent	59	33	92
Column total	160	154	314

Table2The before-and-after sample observations matrix [17 (op cit)]

Populating the McNemar equation with this data results in it appearing thus,

$$X^2 = \frac{(|121-59|-0.5)^2}{121+59}$$

In this example, the null hypothesis of marginal homogeneity would show that the drug treatment had no positive effect. Placing the values in the

McNemar equation results in it generating a value of 21.01, which is not the expected outcome from the distribution of values implied by the null hypothesis. The test therefore, illustrates strong evidence to reach the conclusion that null hypothesis of no treatment effect should be rejected. This test, given here as an example of how non parametric statistics deals with populations of sample data where difference is measured for significance, illustrates the potential for using alternative methods when considering state change over time, rather than for a snapshot of available discrete value data.

Taylor [16 (op cit)] sets out the CPA equations and ascribes the conventional analytical component terms used with this approach. He defines cumulative sum charts (CUSUM) and so-called bootstrapping [19] as the primary inputs to operate on the data. A useful further reference to this concept is described by Hinkley et al [20] where the authors work through numerous examples of what they describe as a mean-shift model, which has the effect of bringing even greater precision to the state change observations by moving the average to compare with every realization of the event sample. To supplement this approach the CUSUM method can be used. This technique from statistical quality control is a sequential analysis technique [15] where it is typically used for monitoring state changes detections. The originator of the CUSUM method,E.S. Page [21], described a quality number θ, as being a parameter of the probability distribution and in particular he used it to refer to the mean. He devised CUSUM as a method to determine changes in (the mean), and proposed a criterion for deciding when to take corrective action. CUSUM relates to a sequence of events; it involves the calculation of a cumulative sum [22]. Samples from a process xn are assigned weights ωn, and summed as follows:

$$S_0 = 0$$

$$S_{n+1} = \max(0, S_n + x_n - \omega_n)$$

When the value of S exceeds a certain threshold value then a change in value has been found. The above formula only detects changes in the positive direction. When negative changes need to be found as well, the min operation should be used instead of the max operation. When this occurs a change has been found when the value of S is below the (negative) value of the threshold value. Figure 6 illustrates a sequence of inputs over time with the change-points mapped according to the threshold values used as parameters for the analysis [15 op cit].

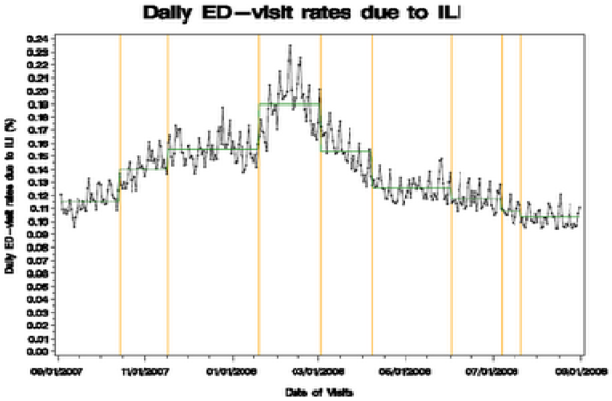


Figure 6 Change Points in a sequence over time

The illustration shown in Figure 6 above is the result of calculating the CPA. Zhiheng [22 op cit] describes the procedure for calculating the CPA in a number of steps thus,

- Determine the Series Mean
- Accumulate Running Sum of differences between Mean and individual values
- Plot CUSUM series
- The point farthest from 0 denotes a Change-Point (CP)
- Break into two sections at CP:
- Analyze each subseries for additional significant CPs

Bootstrapping provides us with a measure of the CP's significance. Taking a statistical perspective on bootstrapping we could assign measures of accuracy to sample estimates using an algorithm designed for the purpose of measuring the properties of the estimator (say its variance) and generate a distribution for re-sampling purposes thereby, bootstrapping the CP with a matrix of covariance values. This is demonstrated later in the paper using a computer program designed to implement the CPA algorithm we will continue now to describe. If we are to model changes-of-state occurrences in a set of continuous data we first have to establish that a change occurred. This requires that we have a single value (or range of values) and the presence or absence of a condition that we can test the datastream against. We also need to know when the change occurred and to what extent it occurred. Cumulatively we need to know how many changes occurred in a given time series because this knowledge helps determine the pattern and severity (significance) of the changes. If we are to use the state-change information as input to another process or decision-making framework we need to know precisely with what confidence can we say that these changes have occurred in the set.Taylor [10 op cit] demonstrates how the confidence level is calculated by performing a large number of bootstraps and counting the number of bootstraps for which, S0diff is less than Sdiff.

So,

let N be the number of bootstrap samples
performed and let X be the number of bootstraps,

where $S^0_{diff} < S_{diff}$.

The confidence level that a change occurred can
then be expressed as a percentage of the sample
thus,

sum((100*X/N)%)

Typically 90%, or 95% confidence is required before one states that a significant change has been detected. For the five bootstraps derived in this example, the values of S0diff are observed as being 7.0, 14.917, 7.975, 7.938 and 9.15 respectively. All of these values are below Sdiff = 17.74167. Figure 7 shows a histogram of S0diff based on 1000 bootstrap samples. Out of 1,000 bootstraps, 995 had S0diff <Sdiff. This gives a confidence level of 99.5%, derived from the sum ((100 = 995/1000)%). This is strong evidence that a change did in fact occur.

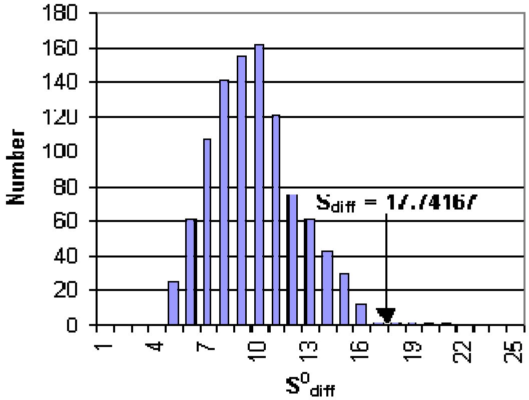


Figure 7Histogram of $S^0_{diff}$for 1000 Bootstrap Samples

A feature of CPA that assists us is that the method features a multiple change detection capability. This is necessary for continuous stream data where the time interval between changes informs us about trends and change magnitudes. The CPA method generates a confidence level and the confidence interval associated with each change that is identified so we can use that information to determine the significance extent of the change and when it occurred.

It is generally held that change-point analysis analytical procedures are extremely flexible. In the examples of how CPA performs with numerous kinds of data, the article [9 (op cit)] demonstrates how it can be used as an analytical tool with various time ordered data types including attribute data, data from non-normal distributions, ill-behaved data such as particle counts and complaint data and data with outliers.

The data

Some 300 million tweet records were collected together for analysis in a single data set. It should be noted that since the year 2009, some 800 million tweets have been collected, which is an indication of the popularity of this social media example. In this study, tweets were filtered according to their location reference for the Japanese cities of Tokyo, Fukuoka and Tokyo. A further filter was that the only tweets analysed were in the English language. For Tokyo, the results depicted in Figure 8 below were obtained.

The tweet results

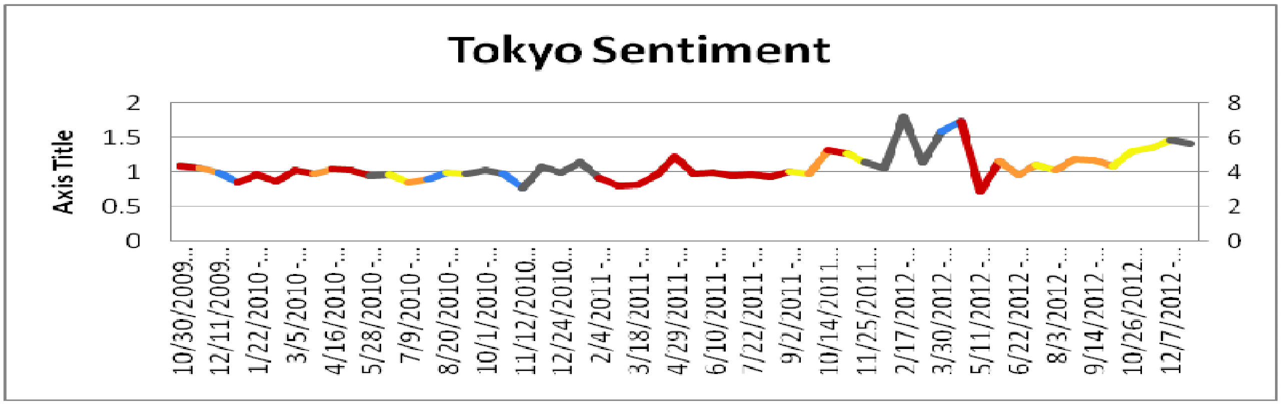


Figure 8– Distribution of tweets relating to Tokyo aggregated over periods of 42 to 45 days. Example tweet text is “In Tokyo Disney Land http://t.co/VhzsjBBV nuclear power, great tohoku earthquake memorial service”. The color indicates the intensity where red means more tweets were collected on this subject and grey indicates fewer tweets were collected on this subject.

Data was culled from tweets on Tokyo from October 2009 to June 2013. We searched on the keyword Tokyo.

We then generated a sentiment measure towards Tokyo using methods decribed in a previous paper [8 op cit] andsought a new method for identifying changes in sentiment using the Change Point Analysis (CPA) algorithm. In doing this, we decided to aggregate the date-sentiment data along the time dimension in 45 day aggregates so the averaged sentiment appears as in Figure 9 below:

Beginning Date of 45 day Period	Average Sentiment per Period
3/05/2010	1.457319279
4/19/2010	1.424474605
6/03/2010	1.44264289
7/18/2010	1.437338075
9/01/2010	1.363525619

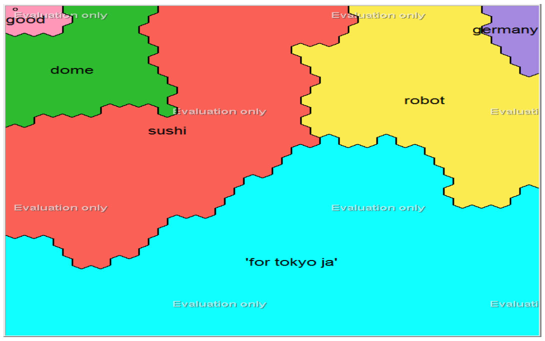
Figure 9 – aggregated data

Three transformations of the sentiment field were observed.

1. Average over the 45 day period.
2. Normalized by dividing average the sentiment by the count of the tweets (plus 1). So,
   normalized value = average sentiment / (total tweet count for the 45 day period +1)
3. Transformed value = 45 day moving average shifted forward by 45 days.

SOM visualizations of tweet conversation

SOM visualizations of tweet "conversation" for each section. For each time section above I did various numbers of clusters of the conversation so we can choose the one which is most outstanding. I guess the main thing to notice is that the conversation is quite "dark" in time section 2 when the tsunami hit.



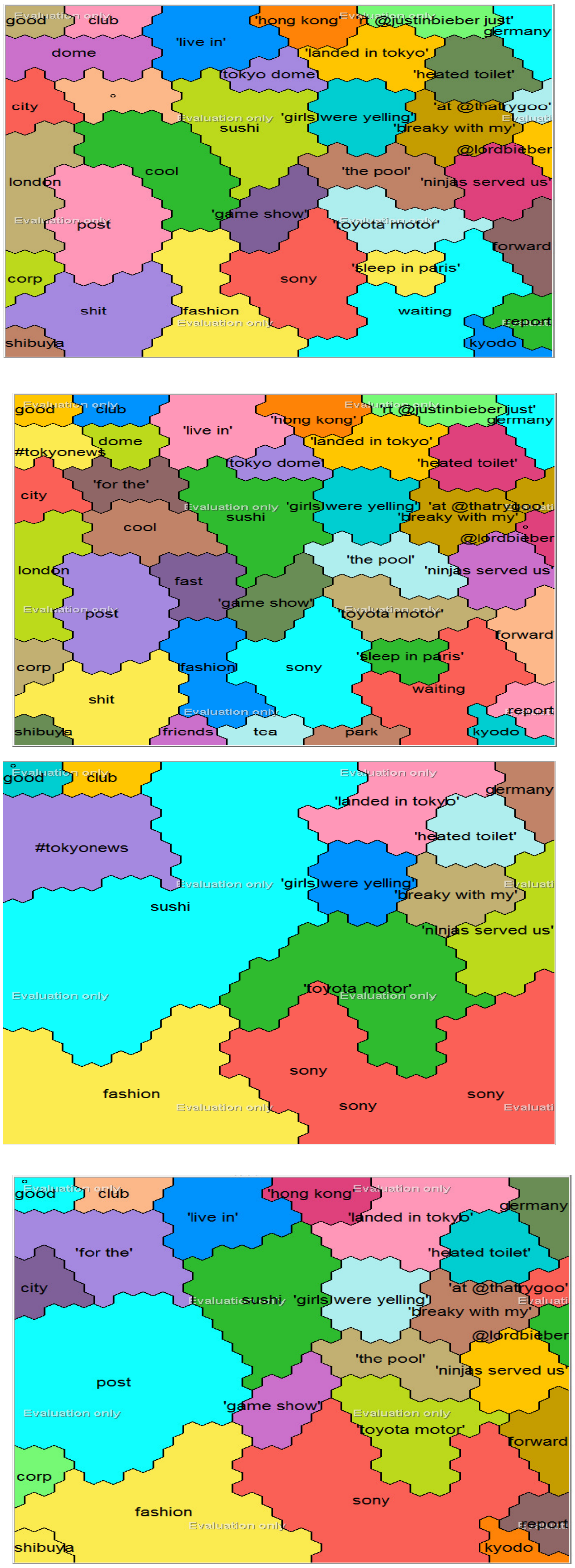
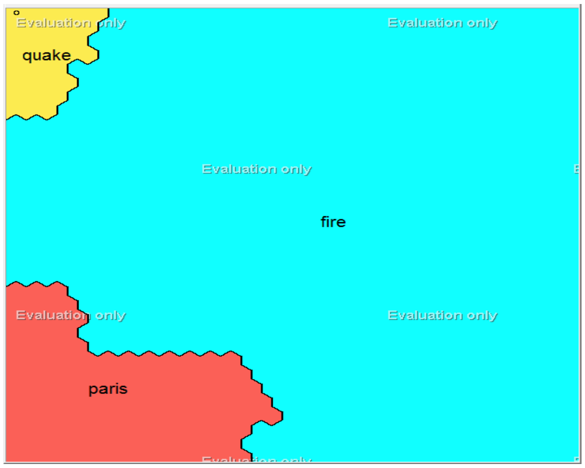
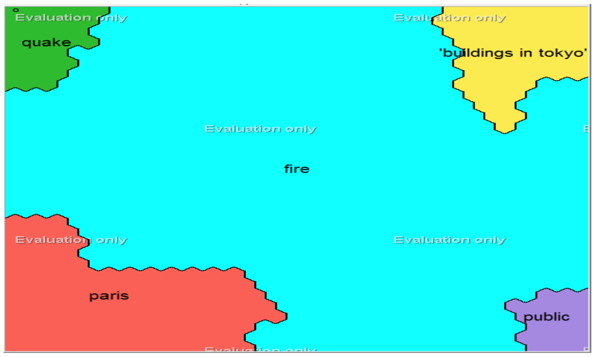


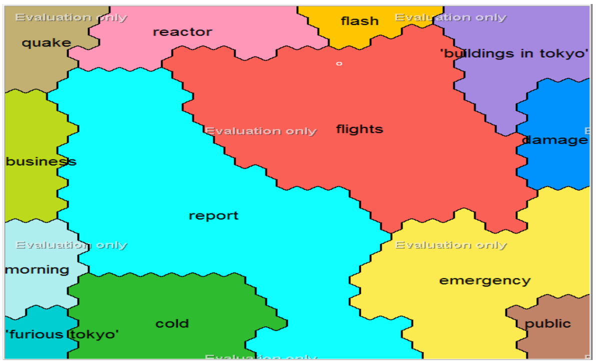
Figure 14: Various views of the conversation taking place in Twitter from October 2009 to October 2010.

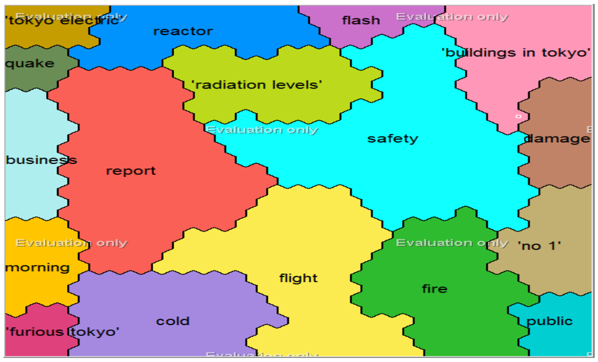
Positive sentiment relating to Tokyo before October 16,2010 was at a relatively high level. From the key words above, it is apparent that many first arrival visitors were for example, impressed amazed by the heated toilets in the airport. Attendance by visitors at live concerts is apparent, with exuberant tweets about the performances, linked with a positive Tokyo experience, evident in the data.

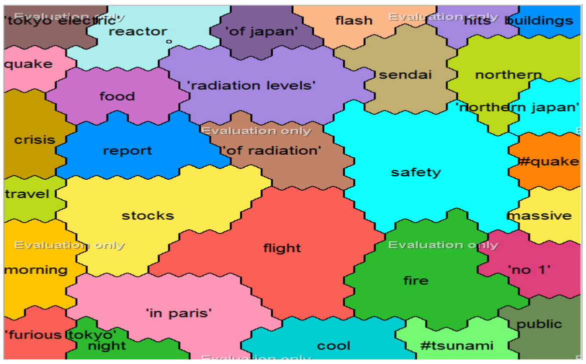
Fashion shopping in Shibuya and positive comments about the purchase of Sony products stand out as tweets in the data but eating Sushi seems to be the greatest joy for the majority of the visitors when they visit Tokyo. The word sushi is connected with many other words in the tweets. This generalised (informal) interpretation of the data reflects a primary useful property of the SOM for tweet analysis – the tool preserves high dimensional topological properties in a 2dimensional mapping. In this way, clusters that are proximate to each other in the map may be related.











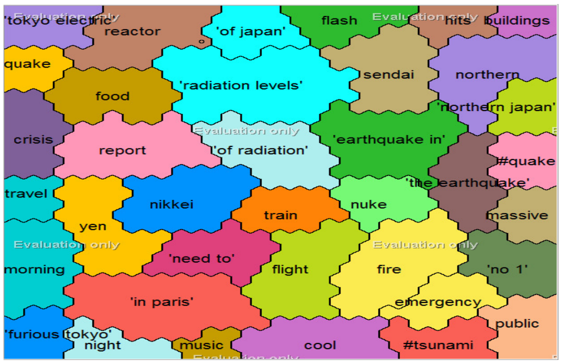
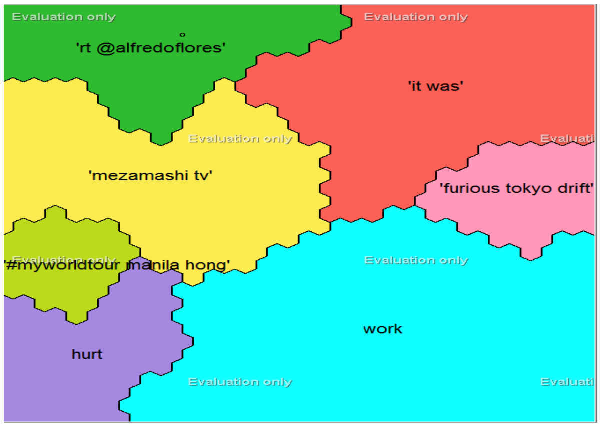
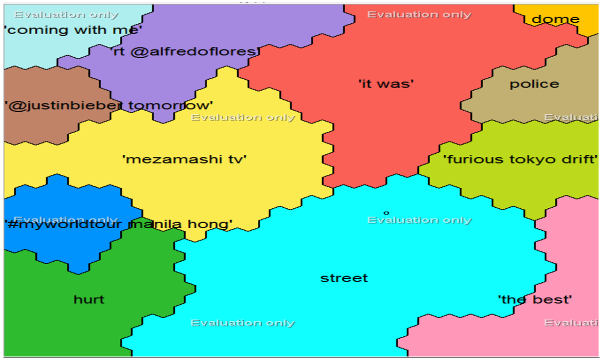
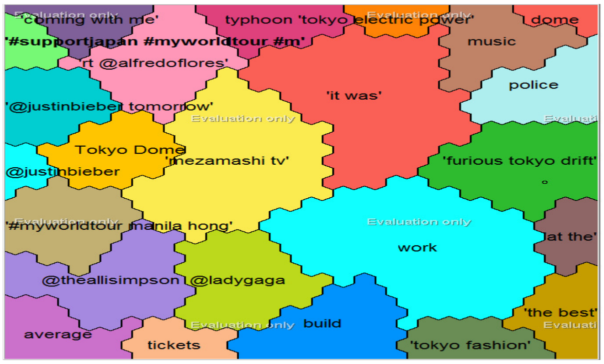


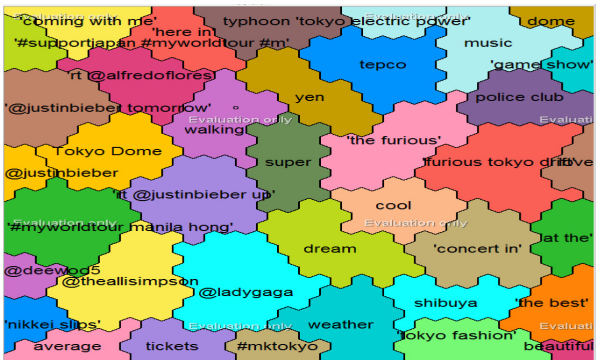
Figure 15: Various views of the conversation taking place in Twitter from October 2010 to April 2011.

In this period, the level of positive sentiment tweets relating to Tokyo dropped fast, mainly due we imagine, to the earthquake and the accompanying worries about radiation and safety. Because of the various emergencies, many flights and trains were cancelled and delayed, which created major travel difficulties for commuters and business travellers. Fire was the main concern at first, thenwhen the Fukushima Power Plant incident occurred, twitter traffic reflected concernsabouttthe extent to which the fallout would effect Tokyo. Moreover, occasional tsunami alerts created a sense of panic, which is reflected in the tweet data. The Japanese economy suffered during this period, with stock values falling dramatically as illustrated by the Nikkei index at the time. Again, this is reflected in the tweet data.









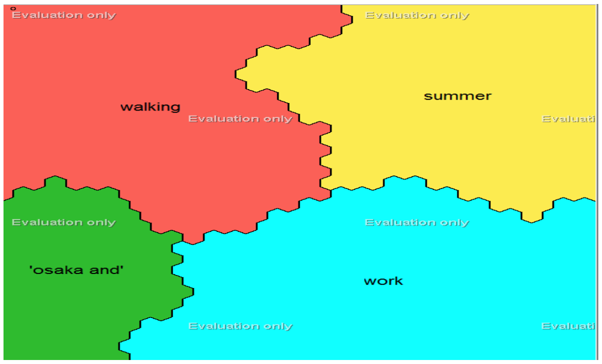


Figure 16: Various views of the conversation taking place in Twitter from May 2011 to March 2012.

In this period, a live concert (JustinBieber) was very popular people in Tokyo. It appears that people have as it were, walked out of the shadow of the earthquake and begun to enjoy the daily life; morning news, work and walk became high use (trend) words on Twitter.

Utility –Can we find improvements in forecasting the Consumer Price Index?

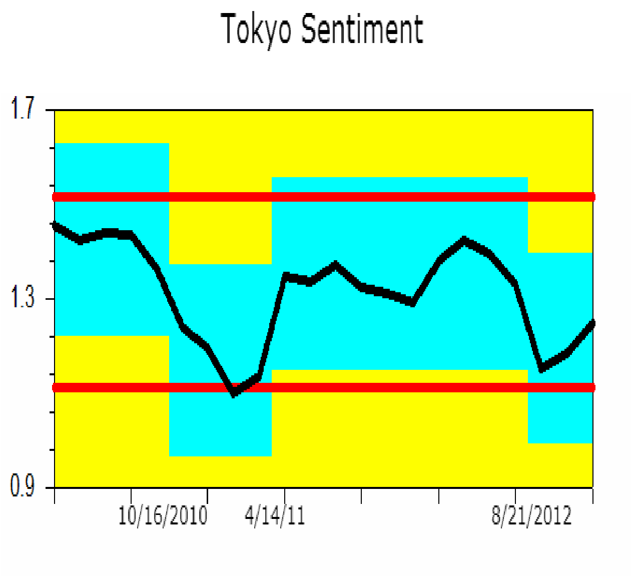


Figure 10: Tokyo sentiment with change points dates.

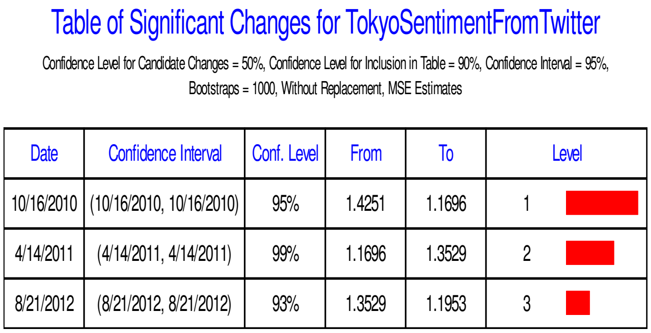


Table of Significant Changes for TokyoSentimentFromTwitter

Confidence Level for Candidate Changes = 50%, Confidence Level for Inclusion in Table = 90%, Confidence Interval = 95%, Bootstraps = 1000, Without Replacement, MSE Estimates

Date	Confidence Interval	Conf. Level	From	To	Level
10/16/2010	(10/16/2010, 10/16/2010)	95%	1.4251	1.1696	1 🟥
4/14/2011	(4/14/2011, 4/14/2011)	99%	1.1696	1.3529	2 🟥
8/21/2012	(8/21/2012, 8/21/2012)	93%	1.3529	1.1953	3 🟥

Table 3

The Change Point Analysis

The analysis process detects three changes. The first change is estimated to have occurred on or about June 1987. This date represents the first month following the change. The second change is estimated to have occurred around Oct 16, 2010. Associated with each change is a confidence level indicating the strength of the change that occurred. The first change occurred with 95% confidence. The second change occurred with 99% confidence. Also associated with each change is a confidence interval for the time of the change indicating how well the time of the change has been pinpointed. 95% confidence is used for all confidence intervals. With 95% confidence, however each of the changes for our data and analysis is on a specific single day. Table 3 also gives additional information about each change. The table indicates that prior to the first change the average sentiment was 1.4251 while after the first change it was 1.1696. At this point it should be pointed out that the sentiment measure is not an absolute measure and actual values are not important but rather changes in sentiment are the useful quantities. Table 3 also gives a level associated with each change. The level is an indication of the importance of the change. The level 1 change is the first change detected and that which is most visibly apparent in the plot in Figure 10. Level 2 changes are detected on a second pass through the data. Any number of levels can exist dependent on the number of changes found.

In order to evaluate and demonstrate the utility of the above described sentiment and change point analysis, tests were conducted to investigate whether the introduction of the sentiment time series and the change point analysis generated intervention variables might improve forecasts of CPI. The test was conducted based on the speculation that sentiment may be used as a supplementto the more laboriously collected consumer confidence data. Innumerable studies have been conducted to forecast CPI including several that include measures of consumer confidence. As recently as 2014, Tsuchiafound that consumer sentiment is not useful for predicting an increase/decrease in household consumption, durable goods consumption, and CPI.

The Data

Economic data was obtained for the Cabinet Office of the Japan Bureau of Statistics and OECD statistics page . Monthly data from October 2010 to December 2012 was included. The data was indexed by setting each attributes value to 100 on the month following the Great Tohoku Tsunami and Earthquake –April 2011.

Prior to indexing, the original variables were: Employment Rate (indexed symbol is IER below), OECD Composite Leading Indicators (indexed symbol below is ILI) from, Consumer Confidence from the Japan Consumer Confidence survey (indexed symbol below is ICC), and Consumer Price Index: total of all items for Japan -seasonally adjusted (indexed symbol below is ICPI). In addition we used our sentiment data as described above. We augmented this data with the construction of three intervention variables based on the change point analysis described above. Each of these was a binary variables with 0 used up until the change of state and 1 thereafter as described next.

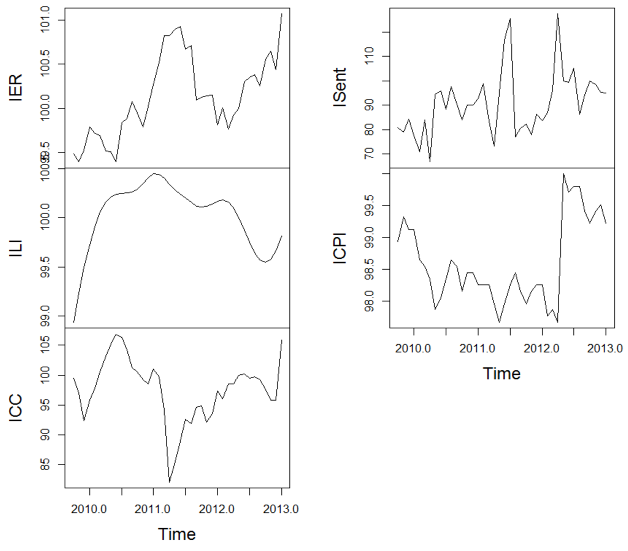


Figure 17: Plots of the five indexed time series variables: Employment, Composite Leading Indicator,Consumer Sentiment, Twitter Sentiment, and CPI.

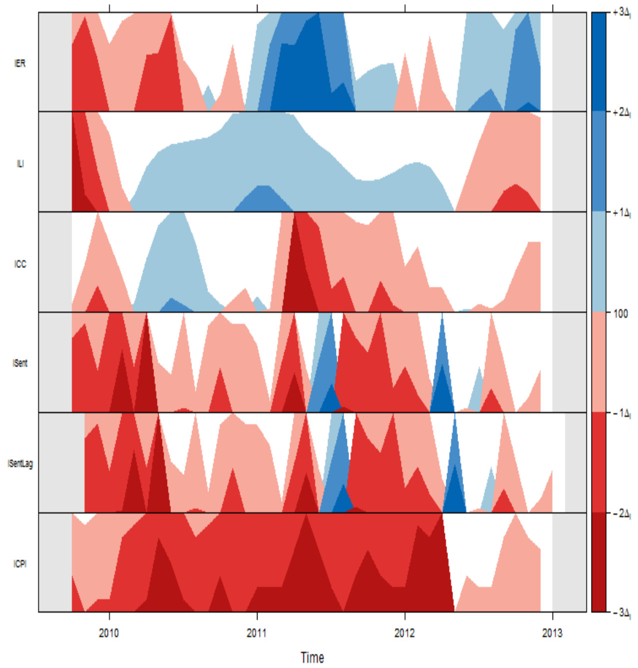


Fig. 18. Horizon plotof indexed explanatory variables for Employment (ER), Composite Leading Indicator (ILI), Consumer Sentiment (ICC), Twitter Sentiment (ISent), one month lagged Twitter Sentiment (ISentLag), and response variable –Consumer Price Index (ICPI). The horizon plot can be used to identify visually leading variables and correlations. It suggests that lagged sentiment (ISentLag) correlates with Consumer Price Index(ICPI).

The Change Point Intervention Variable Methodology (CPIV Methodology)

The objective here is to *make use of the changes of state* identified by the change point analysis described above *for predictive and modeling purposes* so as to extract more value from the sentiment time series. We apply this methodology specifically to our sentiment time series but with almost no modification it can be applied to any time series. For each time t_c, where a change of state occurs define a variable IV_{t_c}which is 1

from $t_c$onwards and before t_c takes on a monotonically increasing sequence of values from 0 to 1. More formally,

$A = \{t|t \geq t_c\}$ and $B = \{t|\ t < t_c\}$

$$\mathrm{IV}_{t_c} = \begin{cases} IV_{t_c}(\mathrm{t}) \ = \ 1 \text{ for t } \in \mathrm{A} \\ \mathrm{IV}_{t_c}(\mathrm{t}) = \ \mathrm{m(t)} \text{ for t } \in \mathrm{B} \end{cases}$$

$where\ m(t)\ is\ monotonically\ increasing\ with\ Range = [0,1]$

For the purpose of our initial study we choose the following special case but it may be worth experimenting less restrictive implementations in future work.

For each time tc, where a change of state occurs define a binary variable?IV?_(t_c) which is 0 before time t and 1 from time t onwards. More formally:

$If\ A = \{t|t\ \geq\ t_c\}\ then\ IV_{t_c} = 1_A(t)\ where\ 1_A(t)\ is\ an\ indicator\ function$

In our example we limited the number of changes of state to a modest number of 3 but the change of state algorithms in various R packages can be tuned to locate a more dense set of changes of state.

Tests

We tested to see if the inclusion of our sentiment measure would improve the forecasts and next whether the inclusion of the change point analysis intervention variables would improve the forecasts. Tests were done with multivariate linear regression and with ARIMA.

Multivariate Linear Regression Models

Four multivariate linear regression models; L1, L2, L3, and L4; as show in table 4 were developed to test whether the inclusion of sentiment and/or the inclusion of the IV variables (described in the prior section) could bring about an improvement in the forecast of CPI. With regard to this regression analysis and using Adjusted R Square as the measure of accuracy we found that inclusion of ISentdid not improve the forecast of the monthly Japanese CPI in the period from 2009 to 2013.However including LaggedISent, a lag of one period in the sentiment measure (as suggested by the horizon plot in Fig. 18) resulted in about a 13 percent increase in the adjusted R squared (see Table 4, Regression results). A further increase in adjusted R squared of about 2.5 percent was obtained with the inclusion of the IV variables. AIC also improved from the L1 model to the L4 model as can be seen in the table. These results are preliminary and further analysis should be conducted but the increased adjusted R square with the addition of the intervention variables derived from the change points suggest that the method may have promise.

Model Name	Explanatory Variables	Multiple R	R Square	Adjusted R Square	AIC	*F*	*Significance F*
L1	IER,ILI, ICC	0.77	0.596	0.549	51.54	18.172	2.39E-07
L2	IER,ILI,ICC, ISent	0.77	0.596	0.549	53.47	13.5413	9.08575E-07
L3	IER,ILI,ICC, ISent, LaggedISent	0.85	0.72	0.68	40.89	16.5658	4.52E-08

L4	Oct10IV, May11IV, Sept12IV[1], IER, ILI, ICC, ISent, LaggedISent	0.876701	0.768604	0.70477	39.83	12.04077	2.18E-07

Table 4. Regression results

Model	L1	L2	L3	L4
Durbin-Watson	1.255	1.2261	1.1911	1.4221
p-value	0.00193	0.00113	0.0007572	0.001848

Table: 5 Durbin-Watson and VIF statistics

Table 5 shows the results of the Durbin-Watson statistic to test for autocorrelation in the residuals for each model. It shows that L4 has the least autocorrelation although all models exhibit autocorrelation. Additionally the VIF factor (variance inflation factor) increased from the model without the intervention variables to the expanded model as shown here:

Variance Inflation Factor (VIF) for L1 model				
IER	ILI	ICC	ISent	ISentLag1
1.838650	1.015003	1.539453	1.243594	1.393218

Variance Inflation Factor (VIF) for L4 model							
IER	ILI	ICC	ISent	ISentLag1	Oct10IV	May11IV	Sept12IV
3.374682	2.690913	2.315456	1.395789	1.535591	3.618339	2.838313	3.139226

This indicates additional multicolinearity in the attributes with the expanded model. The Durbin-Watson statistic and the Variance Inflation Factor (VIF) are utilized to test for autocorrelation in the residuals and multicollinearity in the predictors, respectively. If the residuals are significantly autocorrelated, then the estimated regression coefficients may be inefficient [Wilks, 1995]. Neter et al. [1990] suggest VIF values in excess of 10.0 could cause detrimental effects on the regression coefficients. In all models the VIF was less than 4.

ARIMA Models

Since Durbin-Watson shows some autocorrelation we also tested ARIMA models where the ARMA model is given by

$$X[t] = a[1]X[t-1] + \ldots + a[p]X[t-p] + e[t] + b[1]e[t-1] + \ldots + b[q]e[t-q]$$

We employed parameter search algorithms [Hyndman R. J., Khandakar Y., 2008] to locate the optimal values for the order, integrative, and moving average parameters, searching each parameter between 0 and 12. The best model(see table 6) using the Akaike information criterion (AIC) [Akaike H., 1974] was:

$$X[t] = 1.5115*ar1 + -1.6546*ar2 + 0.8047*ar3 + -0.4719*ar4 + -2.9281*ma1 + 2.9195*ma2 + -0.9912*ma3 + 0.5982*Oct10IV + -0.5418*May11IV + -0.0015*Sept12IV + 0.0286*IER + -1.9867*ILI + 0.0374*ICC + 0.0055*ISent + 0.0334*ISentLag1 + 287.1864$$

The results agree with the multivariate regression results, in that the inclusion of a lagged sentiment variable improved AIC, and that the inclusion of the intervention variables gave a further reduction of AIC. *Moreover* AIC was dramatically reduced from 34.54 to 8.92 when going from the model with the variable set [IER,ILI,ICC, ISent, LaggedISent] (model A1 in table 6) to the full model [Oct10IV, May11IV, Sept12IV , IER, ILI, ICC, ISent, LaggedISent] (model A4 in table 6)and with parameters ARIMA(order=4, integrative =0, moving average =3).

ARIMA (search for optimal model using AIC criteria with p, d, q ranging from 1 to 12)			
Model	Variables	Best ARIMA Model	AIC
A1	IER; ILI; ICC	ARIMA(1,0,0)	47.39
A2	IER; ILI; ICC; ISent	ARIMA(1,0,0)	46.84
A3	IER; ILI; ICC; ISent; ISentLag1	ARIMA(1,0,0)	34.54
A4	Oct10IV; May11IV; Sept12IV; IER; ILI; ICC; ISent; ISentLag1	ARIMA(4,0,3)	8.92

ARIMA (search for optimal model using AIC criteria with p, d, q ranging from 1 to 12)			
Model	Variables	Best ARIMA Model	AIC
A1	IER; ILI; ICC	ARIMA(1,0,0)	47.39
A2	IER; ILI; ICC; ISent	ARIMA(1,0,0)	46.84
A3	IER; ILI; ICC; ISent; ISentLag1	ARIMA(1,0,0)	34.54
A4	Oct10IV; May11IV; Sept12IV; IER; ILI; ICC; ISent; ISentLag1	ARIMA(4,0,3)	8.92

Table 6

Interpretation

It appears that despite issues of collinearity and autocorrelation, the sentiment variables subjected to change point analysis and the therein derived intervention variables(CPIV Methodology)have produced improvements in the accuracy of the multivariate regression models across the accuracy measures Multiple R, R Square, and Adjusted R Square. The increase in the adjusted R squaredwas over 13% from a model without sentiment to one includeing sentiment and a one month lagged sentiment, and a mild increase of 2% with the inclusion of the intervention variables.

The ARIMA methods again confirm that the inclusion of a lagged sentiment variable add value to the model and in fact suggest that it is possible to dramatically improve on AIC with the intervention variable methodology.

Conclusions

We have demonstrated the application of Change Point Analysis techniques as a tool for the interpretation of a very large set of *Twitter Data* expressed as sentiments by individuals. The result from our analysis refelcts a method for gaining accuracy when determining perceptions of reality as represented by this form of social media.The case of locations in Japan, especially sentiments expressed before and after the 2012 earthquake is described and used here to demonstrate the use of Change Point Analysis. The results of the interpretation are proposed as an aid to the continuing quest for finding adequacy in formally modeling for the contemporary phenomenon that is social media and its implications as an economic and business driver. We have shown here that with some innovative use of the sentiment data, it is possible to improve significantly on Consumer Price Index (CPI) forecasts. It appears that this could be a useful and indeed, necessary economic quantity estimator as compared with not utilizing this source of data.

References

[1] Claster, WB. And Cooper, M.Modeling Tourism Sentiment from Twitter Tweets using Naïve Bayes and Unsupervised Artificial Neural Nets

[2] Bollen, J., Mao, H., Zeng, X.Twitter mood predicts the stock market.
Journal of Computational Science, 2011, Elsevier

[3] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe,I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Fourth International AAAI Conference on Weblogs and Social Media.

[4] Fabra, Natalia, and Juan Toro. "The Fall in British Electricity Prices: Market Rules, Market Structure, or Both?." (2003).EconWPA 0309001.

[5] Hillebrand, Eric, and Gunther Schnabl. "The effects of Japanese foreign exchange intervention: GARCH estimation and change point detection." Japan Bank for International Cooperation Discussion Paper 6 (2003).

Andrew, WP., Cranage, DA., Lee, CK. Forecasting Hotel Occupancy Rates with Time Series Models: An Empirical Analysis.Journal of Hospitality & Tourism Research May 1990 (14) 2, 173-182

[6] Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.

[8] Claster, W., Pardo, P., Cooper, M. and Tajeddini, K., (2013) Tourism, travel and tweets: algorithmic text analysis methodologies in tourism, Middle East J. Management, Vol. 1, No. 1, pp. 81-99.

[9] Sallis, P. and Hernandez, S. An event-state depiction algorithm using CPA methods with continuous feed data

[9.5] Rainie L. and Horrigan J. Election 2006 online: Pew Internet & American Life Project Report..

[10]Horrigan J (2008)Online Shopping: Pew Internet & American Life Project Report

[11]ComScore.2007. "Online Consumer-Generated Reviews Have significant Impact on Offline Purchase Behavior," November 29(available online at http://www.comscore.com/).

[12] C. Cardie, C. Farina, T. Bruce, and E. Wagner Using natural language processing to improve eRulemaking. In Proceedings ofDigital Government Research (dg.o), 2006.

[13] N. Kwon, S. Shulman, and E. HovyMultidimensional text analysis for eRulemaking In Proceedings of Digital GovernmentResearch (dg.o), 2006.

[15] Alvarez-Andrade, S. and Bouzebda, S. Some nonparametric tests for change-point detection. DOI: 10.1080/07474946.2014.916930, 12 June 2014 In Taylor and Francis Sequential Analaysis Vol. 33(3), 2014, ISSN 0747-4946 (Print), 1532-4176 (online).

[15] Sallis, P., Claster, W., and Hernandez, S. An algorithm for predicting wind gust events. Computers and the Geosciences, Elsevier, 2011

[16] Taylor, W.Change-point analysis: a powerful new tool for detecting changes. Taylor Enterprises Inc. Libertyville, USA. On-line at http://www.variation.com/cpa/tech/changepoint.html

[17] McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2): 153–157, 1947

[18]A worked example using patient treatment data can be seen at http://en.wikipedia.org/wiki/McNemar's_test

[19]Efron, Bradley and Tibshirani, Robert.An introduction to the Bootstrap. Chapman & Hall, New York, 1993

[20]Hinkley, David and Schechtman, Edna.Conditional bootstrap methods in the mean-shift model. Biometrika, 74 1, 85-93, 1987.

[21] Page, E. S.A test for a change in a parameter occurring at an unknown point. Biometrika, 42, 523-527, 1995.

[22]Zhiheng, X. et al. (2010) Change Point Analysis. On-line at https://sites.google.com/site/changepointanalysis/

[23] Tsuchia. Are Consumer Sentiments Useful in Japan? An Application of a New Market-Timing Test. Applied Economic Letters. DOI:10.1080/13504851.2013.861578Y. 2014

[24] Wilks, D. S., Statistical Methods in the Atmospheric Sciences, Academic Press, New York, 467 pp., 1995.

[25] Neter, J., W. Wasserman, and M. H. Kutner, Applied Linear Statistical Models, Irwin, Burr Ridge, Illinois, 1181 pp., 1990.

[26] Hyndman R. J., Khandakar Y, Automatic Time Series Forecasting: The forecast Package for R, Journal of Statistical Software, July 2008, Volume 27, Issue 3.

[27] Akaike, Hirotugu (1974), "A new look at the statistical model identification", IEEE Transactions on Automatic Control 19 (6): 716–723, doi:10.1109/TAC