Paper 1834-2014

Text Mining Economic Topic Sentiment for Time Series Modeling

Michael P Dessauer, The Dow Chemical Company; Justin Kauhl, Tata Consultancy Services

ABSTRACT

Global businesses must react to daily changes in market conditions over multiple geographies and industries. Consuming reputable daily economic reports assists in understanding these changing conditions, but requires both a significant human time commitment and a subjective assessment of each topic area of interest. To combat these constraints, Dow's Advanced Analytics team has constructed a process to calculate sentence-level topic frequency and sentiment scoring from unstructured economic reports. Daily topic sentiment scores are aggregated to weekly and monthly intervals and used as exogenous variables to model external economic time series data. These models serve to both validate the relationship between our sentiment scoring process and also as near-term forecasts where daily or weekly variables are unavailable. This paper will first describe our process of using SAS® Text Miner to import and discover economic topics and sentiment from unstructured economic reports. The next section describes sentiment variable selection techniques that use SAS/STAT®, SAS/ETS®, and SAS® Enterprise Miner™ to generate similarity measures to economic indices. Our process then uses ARIMAX modeling in SAS® Forecast Studio to create economic index forecasts with topic sentiments. Finally, we show how the sentiment model components are used as a matrix of economic key performance indicators by topic and geography

INTRODUCTION

Financial reports contain a wealth of sentiment on the most up-to-date information regarding economic topics critical to predicting outlooks across industries and geographies. The volume, velocity, and availability of these reports have been increasing over the past several years which creates the opportunity for text analytics to assist the financial analyst in consuming, filtering, and interpreting the data stream. An additional bonus to creating a quantitative assessment of financial reports is to understand the historical view of particular topics in relation to its current sentiment, allowing for assessment in time series forecasting models. If a relationship between an existing structured time series data (Ex: GDP, industrial production, etc..) and relevant sentiment scores, time series forecasting models of external data series could benefit from these near real-time input variables. Our objective is to create a framework which uses daily and weekly economic reports as inputs and generates sentiment time series that both assists in identifying changes in specific topic sentiment and improves forecast accuracy of existing economic time series.

This paper will describe a framework to integrate unstructured sentiment scores into time series forecasting models by using a combination of unstructured data preprocessing, text analytics, statistical analysis, and time series forecasting products. The next four sections will describe each of these phases of the framework in detail, followed by results and conclusions sections. Although the unstructured sources will be kept anonymous for confidentiality purposes, the steps described below to generate sentiment scoring can be replicated and assessed for any source with substantial relevant topic presence and history:

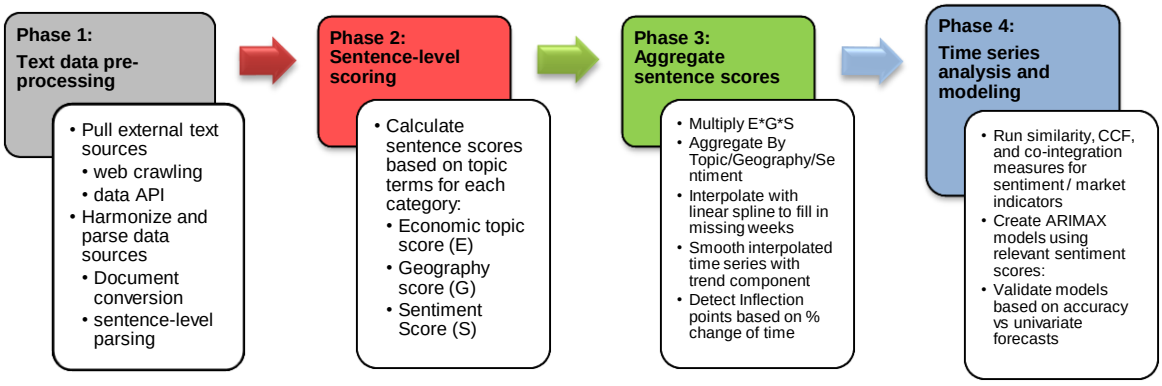


Figure 1. Framework for Sentiment Scoring for Time Series Analysis

1. Text data preprocessing

   Significant data preprocessing is required to convert text from disparate sources into sentence-level samples. Text format conversion, web crawling, and text parsing processes are required to properly condition the text fields prior to analysis.

2. Text mining cleansed data

   Natural language processing and statistical analysis methods convert single-sentence text fields into document-term matrices. Terms are first reduced to include only relevant terms selected by subject matter experts. Topic and sentiment scoring is then performed on each sentence.

3. Topic / sentiment post-processing

   Topic, sentiment, and geographic weights are aggregated and normalized at multiple time frequencies of interest. Once aggregation is complete, the scores are interpolated and smoothed using PROC EXPAND and presented to the subject matter experts for review.

4. Time series analysis of sentiment scores

   Multiple versions of the processed topic/geography/sentiment scores are analyzed and compared to a target series that has a clear relationship to the topic/geography. We use the approach outlined in [1] to determine which version of the sentiment variable to select for modeling.

TEXT DATA PREPROCESSING

Prior to starting any analytics project, it is important to ensure that there is a clean and reliable source of data upon which to base the constructed models. This point still holds true when the basis of those models is unstructured rather than structured. However, greater up front efforts must be expended to transform the raw data into a useable form. SAS Text Miner is tolerant of many kinds of input, but because the system accepts any text-based data that is fed to it does not mean that out of the box results will be reliable or even useful. Considerable manipulation of source data may be necessary to begin receiving reliable results.

Often the modeler will be faced with the task of having to gather external sources of data, which introduces the problems of locating and gathering before any further work can be done. As with any analytics project, copious research into the quality of potential data sources at this stage of the process will help to prevent complications later. Verifying the quality of the data that is being retrieved is necessary for determining how much effort will need to be put into cleaning this data and is a rough indication of the quality of results that you will receive.

The second assertion, after the paraphrased, “good data makes good results”, is that not all unstructured data is created equally; some unstructured data is simply too unstructured or otherwise unsuitable for the intended purpose. Unsurprisingly, the validity of the data derived from unstructured sources is most often directly linked to the quality and professionalism of those sources. Thus it follows that reports by accredited institutions will yield higher returns than information scraped from social media outlets. In most cases, it is relatively easy to spot errors and discrepancies in a numerical data feed. Simply take a count of missing data points or plot the series and patterns in the data will become readily apparent. SAS even provides a number of tools specifically designed to assist the user in detecting potential problems in their structured data sources.

Unfortunately, a large suite of tools for detecting data anomalies is a luxury that is not available when working with unstructured data. SAS provides a few tools in its Text Miner package that allow you to manipulate the data at a granular level, but these will only be useful after you have retrieved and are committed to using a particular set of data. As an extreme example, it is possible to completely randomize the order of words and then the order of letters within those words in a given set of documents and then feed those into SAS Text Miner without a single red flag being raised. This is not a design flaw; it is merely that SAS Text Miner is concerned with different aspects of the text analytics process. The actual cleansing and validating the data prior to being introduced to the SAS system is the responsibility of the modeler. This is why an extensive set of data preprocessing steps was developed for this project, Figure 2.

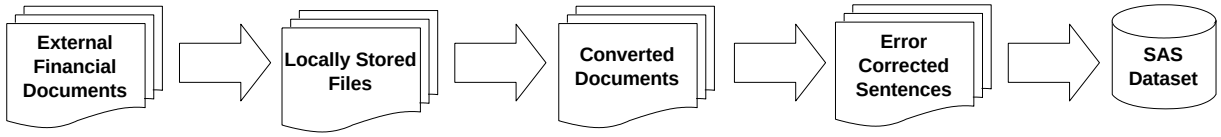


Figure 2. Data Preprocessing

GATHERING DATA

The primary external sources of data for this project were various online suppliers of economic data reports. These sources provide economic reports at various frequencies (daily, weekly, etc) in natural language text, the bulk of which were provided in PDF format. These reports have the advantage of being written by subject matter experts in the field of interest, professionally edited for quality prior to being published online and reliably available for a period of history that was suitable for the later time series modeling.

These sources fulfilled most of the quality conditions that were required, having few upfront errors; the issue then became how to effectively gather this data into a format that could be effectively utilized. One of the most common problems encountered with unstructured data is that useful information within it tends to be fairly sparse and so a large volume of it needs to be collected. For just one of the target sources, several thousand PDF documents were available; the time investment to manually download each would have been prohibitive and each day the new uploads would need to be downloaded to keep the models current. So, in order to accomplish this task in a timelier and far more efficient manner, a web crawler was built and leveraged to automatically extract the data from the sites in question. This tool would automatically download the data into a local repository which could then be easily reached by other tools in the process.

While the web crawling tool used in this process was custom built for the task, it is not always necessary to go to such lengths. There are various freely available web crawlers or web spider tools that are up to the task. SAS Text Miner also provides a built in web crawler tool in its Text Import node, which, while unsuited for downloading files, provides simple web scraping functions. When dealing with the prospect of having to pull gigabytes of unstructured data from the web the modeler will want to have at least one of these options available to them.

CONVERTING AND CLEANSING

In the previous section it was mentioned that the data sources for this project were provided in several different file formats. This is not an uncommon occurrence in text analytics; valuable text data is often conveyed in more than one format depending on the situation. SAS Text Miner is capable of accepting data in several different formats and converting it into a SAS dataset, however the mechanics of these conversions may not be immediately suitable for the data that you have. For example, the Text Import node of SAS Text Miner is capable of reading text data in Excel format, but only if the data within follows a very rigorous column and row format. Chances are great that the external data source is not building their Excel publications with SAS Text Miner's import requirements in mind, so a more hands on approach is inevitable. For this project, converting all documents to plain text format was deemed to be the best approach. In order to accomplish this, conversions had to be made in order to manipulate all raw data into one universal format which SAS Text Miner could then easily read. The chosen format for this project was plain text or ".txt".

The method used for converting a document into plain text depends largely on what format the document is being converted from. For Microsoft Office format files, it is usually best to build the conversion process in Visual Basic or Cscript. For structured documents, like HTML format files, any programming language should be sufficient to decode the highly structured internal format. PDFs can be converted automatically using custom libraries in most programming languages, or the SAS Text Miner Text Import node has this functionality. The emphasis for this stage of the project was on full control of the conversion process. All of the conversions were done using custom built tools where direct manipulation of the variables controlling the conversion process was possible, ensuring that there was minimal damage done to the data in the process.

Once all data was converted into uniform plain text format; another critical concern needed to be addressed, the matter of context. In order to successfully link sentiment with a topic in the later stages of the process a method for limiting the context to which a particular sentiment term applied was necessary. When SAS Text Miner receives an input document it treats all words within that document as appearing in the same context. As most of the economic documents were several pages long and discussed multiple topics along their lengths, a sentiment term which appears in the beginning of the document would be considered to be in the context of a noun phrase which occurs at the end; this is unacceptable and would muddy the results. So, a simplistic process was derived to mitigate this problem. With the assistance of statistical classification on sentence end delimiters; ".", "!", and "?"; it was possible to break the full documents into smaller sentence level documents [2]. The logic in this being that if a sentiment term and a noun phrase both appeared in the context of a sentence, it is highly likely that they will have some kind of relationship.

Once these sentence level documents are formed a final pre-processing step is performed to cleanse the data. It is not uncommon, when dealing with certain high level file formats; to have graphics, diagrams, or other artifacts not useful in the text analysis; embedded into the documents. It is difficult to prevent these from being translated imperfectly into text during the conversion process. These low information content sections of data are segmented out into their own sentence documents fairly reliably by the sentence splitting algorithm, but the problem is in locating

them amidst the other fragments of text. With potentially hundreds of thousands to millions of sentence level documents created by the sentence segmenting tool, sorting by hand is impractical, so two methods were devised to filter out these errant documents.

The first involves simply counting alpha characters against non-alpha characters. Documents which have a high non-alpha character ratio as compared to alpha are a good sign that they were pulled from a non-text artifact, most usually a badly converted chart or graph. While much of this information is useful in other processes, it needed to be removed for text analytics. Other simple mechanical processes were useful as well, such as detecting excessively long strings of text without whitespace and sentences that went on for an excessive length without a clear sentence delimiter.

The second process is more complex. By this point the assumption is that the bulk of the easily detectable nonsense sentences are removed. The target then becomes detecting sentences that appear to have the proper structure, but make no grammatical sense. To accomplish this, the Stanford Parser's Java implemented tool was leveraged to parse each sentence individually [3]. While not one hundred percent reliable, if a sentence was able to parse, then it was deemed acceptable, if not then it was deemed unusable and dropped from the rest of the process.

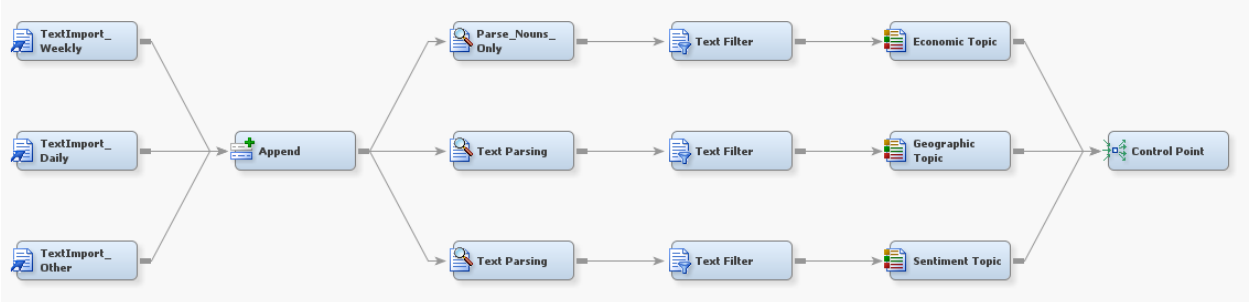
After pushing the raw gathered data through all of these preprocessing tasks, it was then acceptable to load the documents into SAS Text Miner, using the Text Import node, with a measure of assurance that the data would be informative (Display 1). As text data is far noisier and sparse than most other forms of data, extensive preprocessing procedures like these and others are necessary. While time and labor intensive, these processes proved to greatly cut down on the amount of noise being produced by the document sets and, as a result, greatly improved the derived statistics.

	TEXT	uri	NAME	FILTERED	LANGUAGE	CREATED	ACCESSED	MODIFIED	TRUNCATED	OMITTED	EXTENSION	SIZE	FILTEREDSIZE
1	in addition based on the richmond f...	file://E:\projects\...	2007-01-23_1.txt	e:\projects\Text_...	English	08Oct13:12:39:04	07Feb14:21:32:23	08Oct13:12:39:04	0	0	.txt	288	288
2	INTERNATIONAL STRATEGY & I...	file://E:\projects\...	2007-01-24_1.txt	e:\projects\Text_...	English	08Oct13:12:39:04	07Feb14:21:32:23	08Oct13:12:39:04	0	0	.txt	410	410
3	isis homebuilders survey ticked up t...	file://E:\projects\...	2007-01-24_10.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	210	210
4	2 will subprime problems overwhel...	file://E:\projects\...	2007-01-24_11.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	175	173
5	It declined again this week due...	file://E:\projects\...	2007-01-24_12.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	70	70
6	In addition, based on the Rich...	file://E:\projects\...	2007-01-24_13.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	126	122
7	(from However, that's still up fr...	file://E:\projects\...	2007-01-24_14.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	83	81
8	4 will rate resets impact?	file://E:\projects\...	2007-01-24_15.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	165	165
9	Jan 15.3% Richmond Fed's svc ...	file://E:\projects\...	2007-01-24_16.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	69	67
10	In addition, the latest Beige Boo...	file://E:\projects\...	2007-01-24_17.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	103	99
11	we should know in 6 to 12 months.	file://E:\projects\...	2007-01-24_18.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	500	500
12	Real growth is expected to be r...	file://E:\projects\...	2007-01-24_19.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	63	63
13	Over the past few months, home...	file://E:\projects\...	2007-01-24_2.txt	e:\projects\Text_...	English	08Oct13:12:39:04	07Feb14:21:06:11	08Oct13:12:39:04	0	0	.txt	152	152
14	INSTITUTIONAL INVESTORS CO...	file://E:\projects\...	2007-01-24_20.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	173	171
15	U.S. COINCIDENT INDICATOR div...	file://E:\projects\...	2007-01-24_21.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	209	209
16	coin/lag ratio is in recession te...	file://E:\projects\...	2007-01-24_22.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	50	50
17	The coin/lag ratio incorporates...	file://E:\projects\...	2007-01-24_23.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	48	48
18	2 trucking survey still in weakening...	file://E:\projects\...	2007-01-24_24.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	309	309
19	it declined again this week due to w...	file://E:\projects\...	2007-01-24_25.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	358	358
20	U.K. REAL GDP Y/Y % 2006:4Q 3...	file://E:\projects\...	2007-01-24_26.txt	e:\projects\Text_...	English	08Oct13:12:39:05	07Feb14:21:06:11	08Oct13:12:39:05	0	0	.txt	145	145

Display 1. Output from the Text Import Node

TEXT MINING CLEANSED DATA

Once the raw data has been pushed through preprocessing and into SAS Text Miner using the Text import node, it is possible to effectively leverage the text analytics tools provided in the Text Miner package. The goal in this step of the process is to convert the imported noisy text data into useful statistics that can later be leveraged. Three target metrics are defined, Economic Topic, Sentiment, and Geography, upon which data needs to be gathered. To accomplish this, a SAS Enterprise Miner flow is designed, as shown in Display 2.



Display 2. Text into Statistics

If reading in from raw text files as opposed to a SAS data table, SAS Text Miner provides the Text Import node for

importing documents from a directory, or from online using the web crawler function, into a SAS data set. In this project there were several sources of data pulled and so several Text Import nodes were used to target local data repositories which were tied together using an Append node. This pushed all of the data into a single SAS data table from which the rest of the analysis could be based.

As the goal of this process was to pull out information on three metrics, Economic Topic, Sentiment, and Geography, three separate branches were built in the diagram. Each branch handled one and only one of these metrics. In this way they can be calculated independently of each other. This segregation allowed for the ability to manipulate the input data specifically for each topic of concern.

After importing the data, SAS Text Miner requires a Text Parsing node to begin the text analytics process by building a document by term matrix, Display 3. So, each branch begins with this step. For the Economic Topic branch only nouns were considered, reducing the size of the matrix and removing large sets of terms that would otherwise confuse the process. As Sentiment can be conveyed using many different parts of speech, no reduction based on part of speech was performed, but the sentiment dictionary used in a later step was finite and so it was useful to apply a start list to this stage of the process. This reduced the number of terms considered to only those in the sentiment dictionary. Geographic terms are commonly only expressed as nouns, but as with the Sentiment branch, all topics later in the process were custom, so it was more expedient to simply apply the dictionary as a start list in this step. Even with these filters applied, the resulting matrix is very sparse (Display 3-4).

	TERMNUM	_DOCUMENT_	_COUNT_
1	15	1	1
2	6	1	1
3	8	1	1
4	3	1	1
5	11	1	1
6	1	1	1
7	4	1	1
8	12	1	1
9	17	1	1
10	16	1	1
11	10	1	1
12	7	1	1
13	2	1	1
14	5	1	1
15	13	1	1
16	9	1	1
17	14	1	1
18	29	2	1
19	18	2	1
20	26	2	1

Display 3. Example of Economic Topic parsing matrix

term	Role
accelerate	Verb
accelerate	
accommodative	Adj
advance	
advantage	Noun
affordable	Adj
alarm	Verb
alarming	Adj
amaze	Verb
amazing	Adj
ample	Adj
applaud	Verb
appreciate	Verb
appropriate	Verb
approval	Noun
approve	Verb
astound	Verb
attack	Noun
attack	Verb
attractive	Adj
available	Adj
awful	Adj
bad	Adj

Display 4. Sentiment Start List

SAS Text Miner requires that a Text Filter node always follows a Text Parsing node and so, in each of the three branches, this is the next step. SAS Text Miner automatically filters out low information content words, but, as a stop

list was used in the Text Parsing node of the Sentiment and Geography branches, this node served little purpose other than to convert the matrix into the proper format for the next step. In the Economic Topic branch this node was able to perform its full function by filtering terms which occurred below a specific threshold of times. It was also customized slightly to exclude terms which did occur frequently enough to be useful, but were not particularly useful for this analysis.

From here the actual conversion of text into useful metrics can begin. To accomplish this, the Text Topic node was employed. The benefit of the text topic node in this situation over the Text Cluster node is that it allows for a single document to be aligned to multiple topics whereas the Text Cluster node forces each document to be aligned to one and only one cluster. This is useful, as in many cases the documents will mention a single topic of concern and then multiple regions in which it is a concern or multiple topics of concern are mentioned for a single region. In such cases classifying the document into one and only one bucket would be an incorrect approach.

The analysis for both the Sentiment and Geographic branch Text Topic nodes was fairly straightforward. A set of topics was derived by hand previously and inputted into the node. For the Geography branch, these were terms such as “China” or “Washington, D.C.” which would assist in clearly bucketing the documents into particular geographies of interest. For Sentiment, terms such as “rising” or “uncertain” were used to bucket the documents into either expressing positive or negative sentiment. Some caution needed to be taken at this point as sentiment terms were not always universally valid. For example, “rising GDP” expresses a completely different sentiment than “rising unemployment”.

Defining the topics for the Economic Topic branch was not as easy. One of the goals of this project was to define the topics that should be of interest around these documents. So, to determine these topics a first exploratory pass was performed on the Economic Topic Text Topic node using twenty five automatically generated topics. Examples of these can be viewed in Display 5. These were then reviewed and roughly defined before bringing them before a subject matter expert who checked them for relevancy. The subject matter expert identified those topics which made the most sense while outlining a few more that the tool missed. This information was then applied to create custom topics from future runs. This process was iterative, each time refining the topics until the subject matter expert was pleased with the results.

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
+exposure,+fund,equities,net,max	Multiple	**0.289**	**2.976**	117	1669
bearish,institutional,bullish,min,duration	Multiple	**0.25**	**2.103**	93	1732
institutional bond,+bond,mgrs,sloterbeck,+oscar	Multiple	**0.218**	**1.822**	54	2334
wk,+strong,avg,+weak,+sale	Multiple	**0.212**	**1.727**	111	5864
+market,stock,+stock market,+stock,china	Multiple	**0.163**	**0.891**	150	2388

Display 5. Example of Automatically Generated Economic Topics

Once all three of these branches were completed it was then possible to extract the raw statistical information on any cross-section of the data along any of these three metrics. The data derived from these steps reflects the frequency of occurrence of these topics across the input data corpus. The assumption maintained is that as an economic topic is talked about more frequently, it becomes more relevant in the markets. By using sentiment and geography it is possible to determine the nature of this shift and identify where it will become a relevant concern. This raw numerical data is used as the input for the next set of procedures which deal with preparing it for eventual use in economic models.

TOPIC SENTIMENT POST PROCESSING

The third stage in the project is sentence-level economic topic sentiment by geography conversion into time series, that serves as inputs into KPI visualizations and ARIMAX forecast models. Our process aggregates positive and negative scores into weekly buckets using the economic topic and geography weights (Figure 3). After the scores are aggregated, we use PROC EXPAND to transform the series using a linear spline for interpolation and a trend component for smoothing. KPIs based on changes to direction of transformed were created that allow decision makers to visualize changing sentiment trends over time.

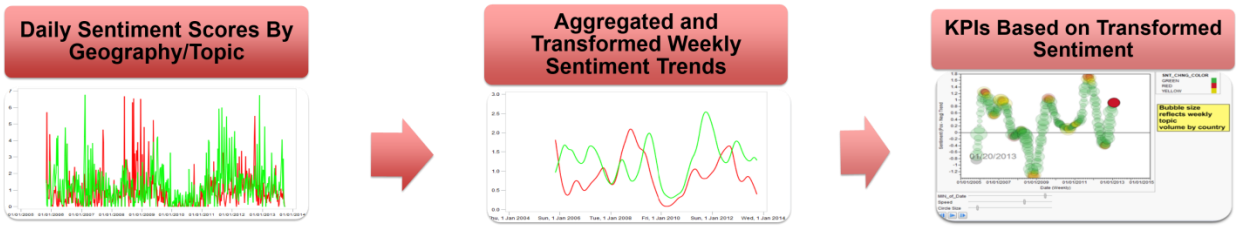


Figure 3. Topic Sentiment Series Examples Showing Transformation from Sentiment to KPI Visualization

RAW SENTIMENT CALCULATION & AGGREGATION

The first post processing step after the text mining stage is to calculate the raw sentiment scores for each topic/sentiment/geography combination over a specified time interval. The aggregated sentiment calculation $\boldsymbol{S_t}$ for time t is calculated using the formula below, where n is the total number of sentences in time t, E_i is the raw economic topic weight, Γ_i is the binary geography weight (0,1), and P_i is the raw sentiment weight.

$$\boldsymbol{S_t} = \sum_{i=1}^{n} \mathrm{E}_i \Gamma_i \mathrm{P}_i$$

A positive and negative sentiment series is generated for every topic/geography combination and reviewed by subject matter experts. Several methods for calculating sentiment were tested, which included binary topic and sentiment calculation as well as normalizing aggregated values by total topic mentions. It was concluded by subject matter expert review, that raw aggregation without normalization provided the most reflective historical view of topic sentiment (Figure 4).

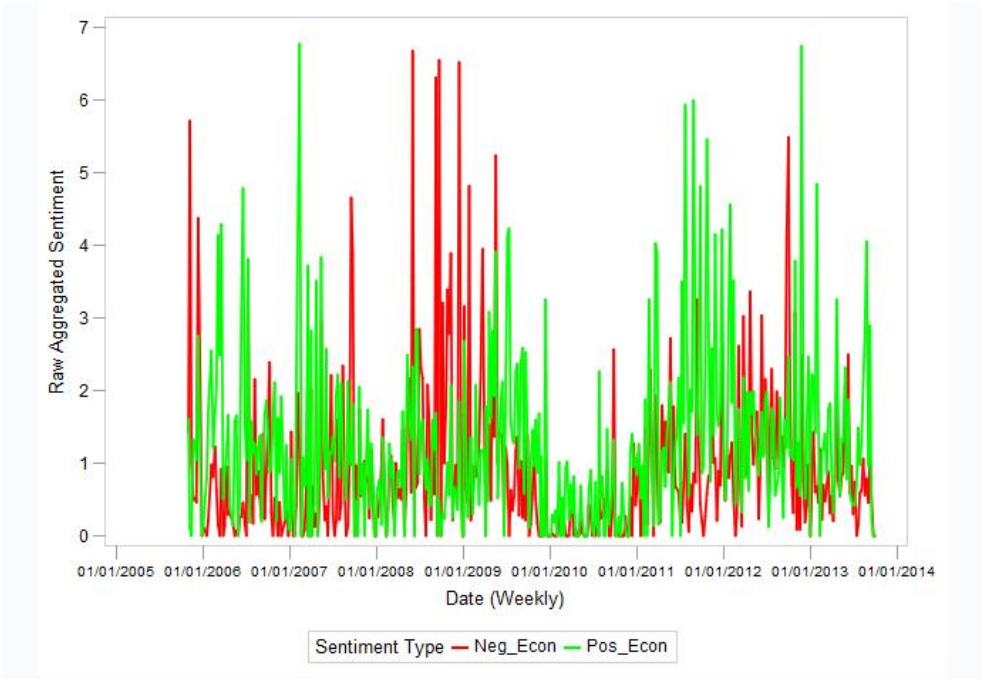


Figure 4. Weekly Aggregated Positive and Negative Topic Sentiment Time Series Example

AGGREGATED SENTIMENT TRANSFORMATION

As seen in the figure above, aggregated sentiment scores(especially high frequencies such as daily or weekly) still contain a significant amount of fluctuations. We perform data transformations to smooth and interpolate the series. We use PROC EXPAND to first interpolate the series followed by the Hodrick-Prescott Filter trend component (code below).

```
PROC EXPAND DATA=WORK.RAWSENTDATA
    OUT=WORK.SMOOTHSENTDATA
```

```
      TO = WEEK
      ALIGN = BEGINNING
      METHOD = JOIN
      OBSERVED = (BEGINNING, BEGINNING);
      BY Sntmnt;
      ID Date;
      CONVERT AGGREGATE_RAW_SENT /
            TRANSFORMOUT =(HP_T 1600);
RUN; QUIT;
```

After the transformation step is complete, we share the results with subject matter experts to review the historical trends. Historical validation with sentiment trend directional changes is initially compared with known events that should align with specific topic/geographies. This manual validation assists in refining transformation and sentiment aggregation calculations. Figure 5 shows how the transformation affects the raw aggregated sentiment series shown in the previous figure.

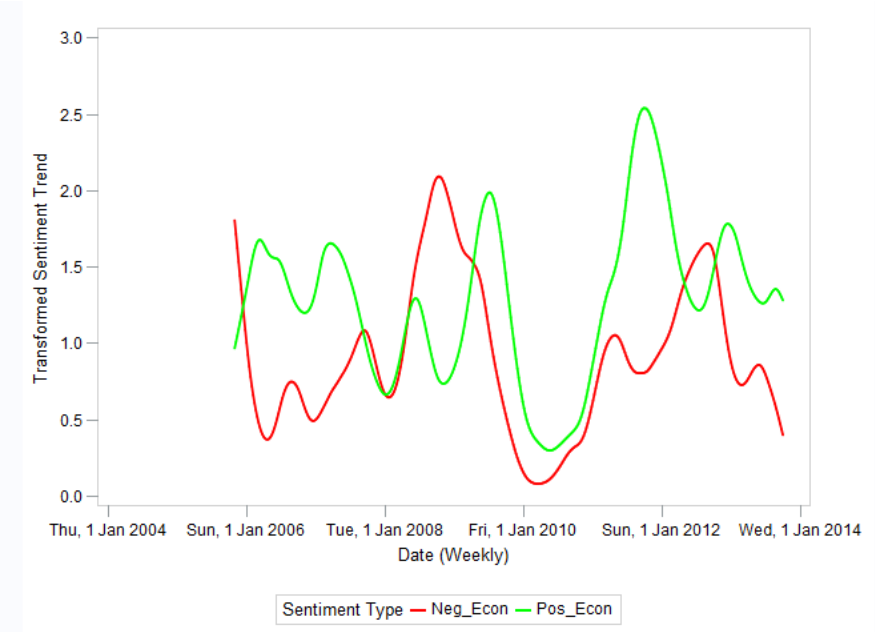


Figure 5. Transformed Weekly Positive and Negative Topic Sentiment Time Series Example

CALCULATE INFLECTION POINTS FOR KPI PLOTS

Now that the data is transformed for use in KPI plots and as inputs into time series analysis steps, we first take the difference between the positive and negative trends at each time point then calculate difference percent change over the previous time point (shown in PROC EXPAND sub-step below).

```
      TRANSFORMOUT =(PCTDIF 1)
```

The intent of the KPI is to identify "inflection points" in the sentiment time series trends so that analysts can quickly focus on potentially changing market conditions. These inflection points occur when the absolute value of the first difference percent change approaches zero. We classify the absolute % change interval into three categories as shown below:

- Green: $|\%diff| > 3\%$
- Yellow: $3\% > |\%diff| > 1\%$
- Red: $|\%diff| < 1\%$

Two types of KPIs are created as our first analysis objective of the project using the sentiment trends and labels described above. The first KPI uses the SAS® JMP™ bubble plot to visualize the changing sentiment (y axis) over

time while identifying the inflection point labels (color) and total volume of topic mentions (bubble size). A sample bubble plot KPI is shown below in Figure 6.

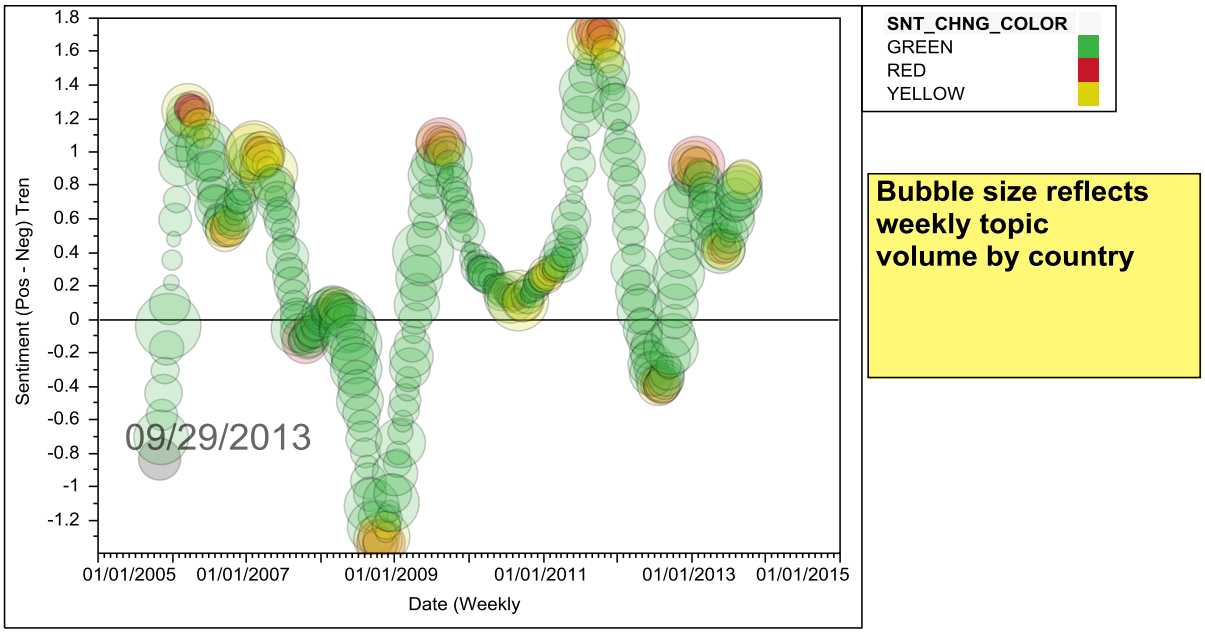


Figure 6. Example of the bubble plot KPI for showing changing sentiment over time

In the situation where an analysts prefers to see all geographies at once for a specific time period, our second KPI gives a snapshot of the sentiment inflection points or general sentiment (Figure 7). We use the map shape feature in the JMP Graph Builder to build global sentiment color maps.

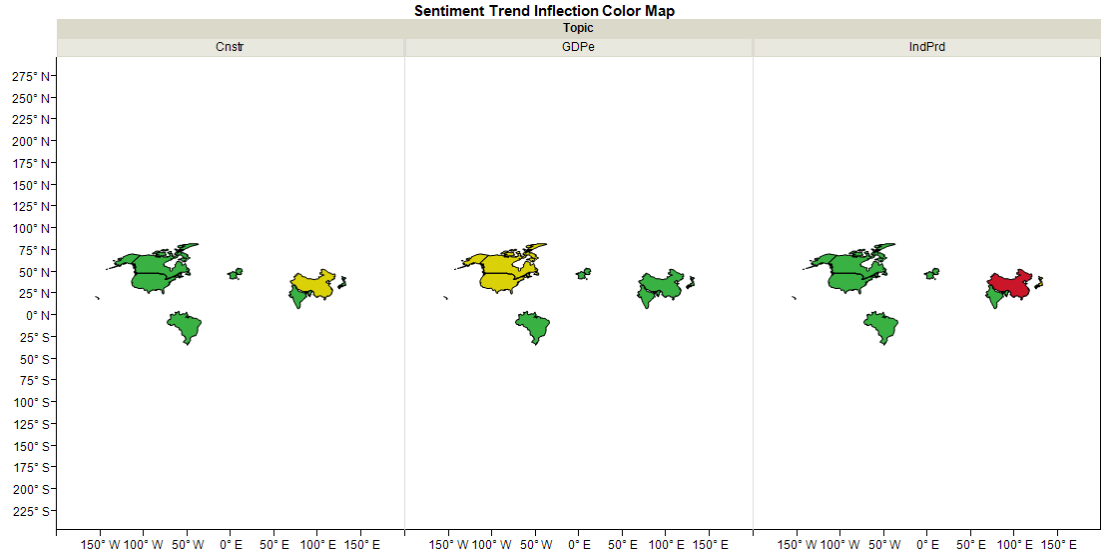


Figure 7. Sentiment Inflection Color Maps – By Topic and Geography

The final phase of the analysis is to use the transformed sentiment time series as inputs to forecast external economic time series. The three series selected as test cases are collected in monthly intervals, thus a time series compression is required for the comparison. We elected to aggregate the weekly sentiment time series to the monthly time frequency by simply summing up the weekly data points over each month (all weeks with start date within the month).

TIME SERIES ANALYSIS USING SENTIMENT SCORES

The final phase of the analysis explores the utility of using transformed sentiment trends as leading indicators for external market indices. The statistical time series comparison to key market variables is also used as a validation step in confirming financial report sentiment has similar behavior to these indices' historical patterns. We first compared the monthly version of the transformed series with external market indicators using the time series analysis processed described in [1] which includes similarity, cross-correlation, and co-integration analysis. After key relationships are identified, time series ARIMAX models are generated via SAS Forecast Studio. We will describe the variable selection analysis and time series modeling steps in more detail in the remainder of this section.

EXTERNAL INDEX TEST CASES

Three different indices were selected as test cases to provide coverage over different geographies, economic topics, and relevant topic volume throughout the economic report histories. The three indices described below:

- **NAHB/Wells Fargo Housing Market Regional Index [4]:** Index SA Source: National Association of Home Builders (NAHB)/First American NOTE: On a scale of 100, 0=worst mkt conditions, 100=best, 50=avg
- **China Industrial Production, seasonally adjusted de-trended [5]** Units: index, base year 2010=100 Data source: China National Bureau of Statistics Data edge: 2013m11 Forecast source: IHS Global Insight Updated monthly. Last update: 15 December 2013
- **US 2011 Annual GDP Forecast - Consensus Economics [6]**

The economic topics selected by our subject matter experts specifically align to these three topic / geography external data series. We then test all relevant sentiment scores to identify relationships between the external indices and sentiment using the methods outlines in the next section.

SENTIMENT VARIABLE SELECTION FOR TIME SERIES MODELING

To identify statistically significant relationships between sentiment time series and the dependent described in the previous section, we employ the variable selection methodology described in [1] which are broken up into three methods:

1. A Similarity analysis approach can be used for both variable reduction and variable selection. Leonard and [7] introduce via PROC SIMILARITY in SAS Econometrics and Time Series (ETS), an approach for analyzing and measuring the similarity of multiple time series variables. Unlike traditional time series modeling for relating Y (target) to an X (input) similarity analysis leverages the fact that the data is ordered. A similarity measure can take various forms but essentially is a metric that measures the distance between the X and Y sequences keeping in mind ordering. Similarity can be used simply to get the similarity between the Y"s and the X"s but it can also be used as input to a Variable Clustering (PROC VARCLUS in SAS STAT) algorithm to then get clusters of X"s to help reduce redundant information in the X"s and thus reduce the number of X"s.
2. A co-integration approach for variable selection is a supervised approach. Engle and Granger [8] discuss a co-integration test. Co-integration is a test of the Economic theory that two variables move together in the long run. The traditional approach to measuring the relationship between Y and X would be to make each series stationary (generally by taking first differences) and then see if they are related using a regression approach. This differencing may result in a loss of information about the long run relationship. Differencing has been shown to be a harsh method for rendering a series stationary. Thus co-integration takes a different tack. First, the simple OLS regression model (called the co-integrating regression), the residual are obtained use the exogenous variables as the Y and the dependent variable as the X. Then, a test statistic is used to see if the residuals of the model are stationary. This test can be either the Dickey-Fuller Test or the Durbin Watson Test. In the implementation examples below the Dickey-Fuller test is used.
3. A cross correlation approach for variable selection is also a supervised approach. A common approach used in time series modeling for understanding the relationship between Y"s and X"s is called the Cross Correlation Function (CCF). A CCF is simply the bar chart of simple Pearson Product moment correlations for each of the lags under study.

Based on a combinatorial assessment of these three approaches, we have identified best relationships between sentiment scores and the external indices to test as input variables in time series ARIMAX modeling. The figures 8-10 below show the sentiment score and external index cross series plot and corresponding CCF plot for the sentiment series with the strongest relationship to the y series.

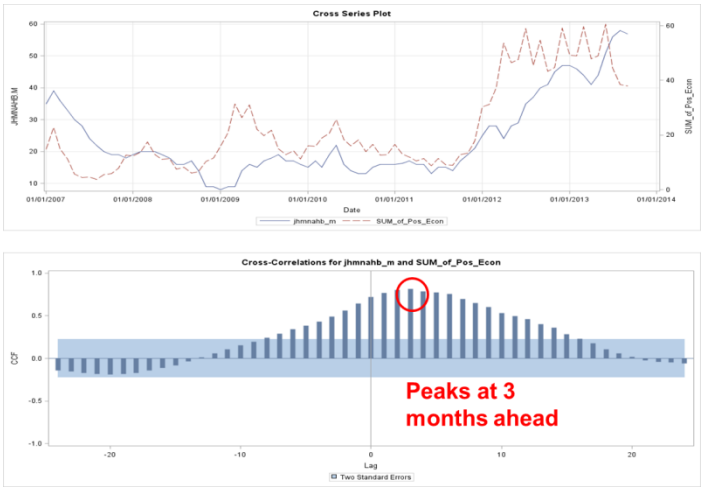


Figure 8. US Construction Positive Sentiment Cross-series plot and CCF plot vs AHB/Wells Fargo Housing Market Regional Index

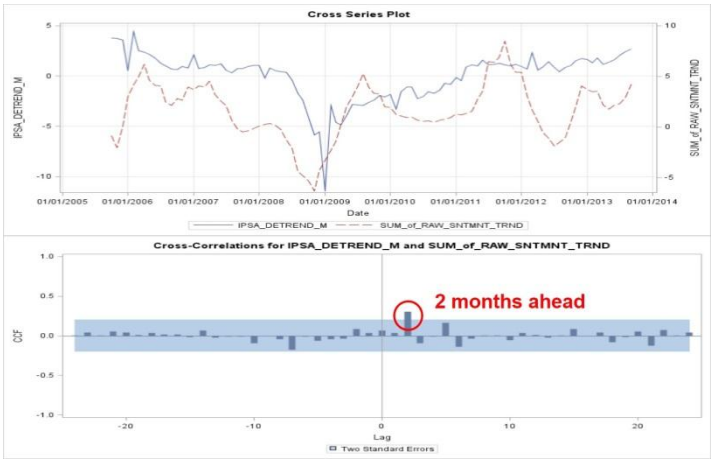


Figure 9. Chinese Industrial Production Sentiment Cross-series plot and CCF plot vs De-trended China Industrial Production Index

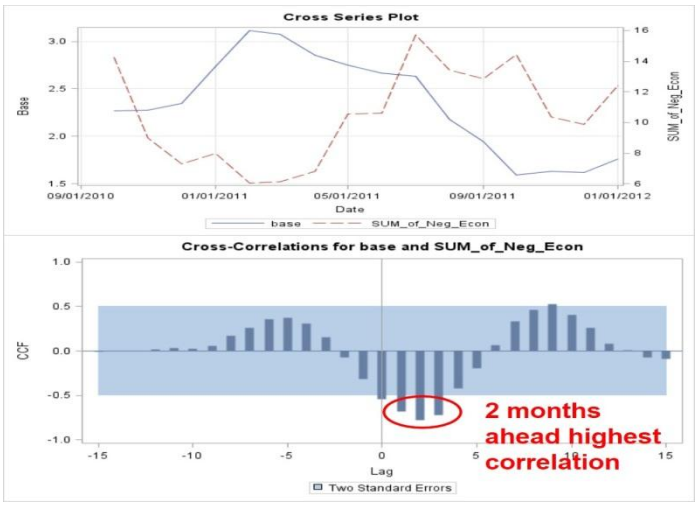


Figure 10. US GDP / Economy Negative Sentiment Cross-series plot and CCF plot vs Consensus Economics 2011 GDP Annual Forecast

ARIMAX MODELING USING SENTIMENT VARIABLES

Now that sentiment input variables have been identified for each of the three external indices, we proceed to create forecast projects using SAS Forecast Studio. Our objective is to create a forecast model that uses the identified sentiment variable as a statistically significant model component. We will then test this model against the best fit univariate model and compare model accuracies over multiple hold out intervals.

US CONSTRUCTION - AHB/WELLS FARGO HOUSING MARKET REGIONAL INDEX MODEL

For the US Construction topic, we create a model using the positive construction topic sentiment trend variable and tested the multiple components. A best model based on the Symmetric Mean Absolute Percent Error (SMAPE) accuracy measure for a single hold out length is shown below:

Component	Parameter	Estimate	Standard Error	t Value	Approx. Pr > |t|
Dependent_Variable	AR1_1	1.23884	0.12171	10.18	<.0001
Dependent_Variable	AR1_2	-0.34537	0.19121	-1.81	0.0751
Dependent_Variable	AR1_3	0.0925	0.12195	0.76	0.4506
Positive_Sentiment	SCALE	0.03619	0.04084	0.89	0.3785
Positive_Sentiment	NUM1_1	-0.07574	0.02609	-2.9	0.0049
Positive_Sentiment	NUM1_2	-0.0056	0.04096	-0.14	0.8916
Positive_Sentiment	DEN1_2	0.94689	0.0209	45.31	<.0001

Display 6. US Housing Market Index Model Sentiment Model Components

Based on the t values, two of the sentiment variable components (Num 1 and Den 2) show significance in the model.

CHINESE INDUSTRIAL PRODUCTION SENTIMENT MODEL

The de-trended Chinese Industrial Production Index model uses the Chinese industrial production topic sentiment trend (positive trend – negative trend) variable. Two components of the sentiment variable are significant in the model (Scale, Den1_1).

Component	Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
DEPENDENT	CONSTANT	2.61979	1.01098	2.59	0.0113
DEPENDENT	AR1_1	0.5045	0.09657	5.22	<.0001
DEPENDENT	AR1_2	0.46807	0.09691	4.83	<.0001
SENTIMENT	SCALE	0.40123	0.13364	3	0.0035
SENTIMENT	NUM1_1	-0.14842	0.13486	-1.1	0.2742
SENTIMENT	NUM1_2	0.33101	0.12801	2.59	0.0114
SENTIMENT	DEN1_1	-0.84939	0.26166	-3.25	0.0017
EVENT_SEPT09_PULSE	SCALE	0.73106	0.73278	1	0.3213

Display 7. China Industrial Production Sentiment Model Components

US 2011 GDP SENTIMENT MODEL

The GDP model uses the negative US GDP/Economy topic sentiment as a leading indicator in forecasting the annual GDP baseline forecast by consensus economics. The negative sentiment components do not show significance based on the t value, but improve model accuracy for the test hold out length.

Component	Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
DEPENDENT	MA1_1	-0.90432	0.32232	-2.81	0.0205
DEPENDENT	MA1_2	-0.744	0.27108	-2.74	0.0227
Negative Sentiment	SCALE	-0.02251	0.02063	-1.09	0.3035
Negative Sentiment	NUM_1	-0.00923	0.007736	-1.19	0.2633

Display 8. Consensus Econmics 2011 GDP Forecast Sentiment Model Components

ARIMAX MODEL VALIDATION

After the models with identified sentiment variables have been created, a univariate forecast model is also generated using the same test hold out. This univariate model is used a baseline accuracy measure to compare the overall additional model accuracy gained by incorporating sentiment variables in forecast models. In each case, we used equally spaced hold out intervals that would also allow model estimation process to use at least 75% of the data points.

MODEL RESULTS

The three model SMAPE accuracy measurements over the defined hold out intervals are provided in this section. An average SMAPE is also included in the tables. Each model measurement shows an improvement in model accuracy when compared to the univariate forecasts.

	AHB/WELLS FARGO HOUSING MARKET REGIONAL INDEX SMAPE By Hold Out Interval (Monthly)						
Model Type	**In-Sample**	**3**	**6**	**12**	**15**	**18**	**Avg**
Univariate	8.55%	2.34%	11.75%	6.92%	29.76%	34.07%	**15.57%**
Sentiment Model	8.13%	2.44%	9.05%	9.10%	11.06%	22.09%	**10.31%**

Display 9. Comparison of US Housing Market Index Model Accuracy between Univariate and Sentiment Models

The overall average SMAPE for the US construction sentiment model across the test hold out intervals and in-sample shows a 33% improvement in model accuracies when compared to the best univariate model.

	China Industrial Production, seasonally adjusted de-trended SMAPE By Hold Out Interval (Monthly)				
Model Type	**In-Sample**	**3**	**6**	**12**	**Avg**
Univariate	78.99%	38.76%	25.52%	57.30%	**50.14%**
Sentiment Model	59.75%	29.01%	48.61%	27.25%	**41.16%**

Display 10. Comparison of Chinese Industrial Production Index Model Accuracy between Univariate and Sentiment Models

The overall average SMAPE for the Chinese Industrial production sentiment model across the test hold out intervals and in-sample shows a 17% improvement in model accuracies when compared to the best univariate model.

	2011 Consensus Economics GDP Forecast SMAPE By Hold Out Interval (Monthly)			
Model Type	**In-Sample**	**1**	**2**	**Avg**
Univariate	6.89%	8.28%	4.14%	**6.44%**
Sentiment	5.50%	0.66%	5.50%	**3.89%**

Display 11. Comparison of Consensus Monthly GDP Forecast Model Accuracy between Univariate and Sentiment Models

The overall average SMAPE for the US GDP/Economy sentiment model across the test hold out intervals and in-sample shows a 40% improvement in model accuracies when compared to the best univariate model.

CONCULSIONS FOR TIME SERIES MODELING USING SENTIMENT TRENDS

After studying both the text mining topic, sentiment, and geography sentence scoring methodology and the time series modeling results, we have concluded the following:

- Financial report sentence-level sentiment scores with sufficient topic / geography volume have shown statistically validated relationship with similar external indicators
- A reduced set of financial sentiment terms can produce strong relationships to external indicators than exhaustive sentiment term sets
- Although ARIMAX models with sentiment score variables demonstrate improvements in model quality, additional leading indicator variables typically improve models.
- Inflection point identifying a change in sentiment trend direction can be used as an early warning signal, although reliable forecasts should still include structured leading indicator variables

These conclusions have given our analytics team enough confidence to pursue improvements in sentiment quality though additional text processing and scoring methods. Our next steps include working at the noun-phrase level to reduce inaccurate topic-sentiment-geography relationships that occur at sentence and creating topic-level sentiment through supervised training methods. We will also increase the taxonomy of topics and geographies to include hierarchical economic topic ontology.

REFERENCES

[1] Timothy Rey, Chip Wells, and Justin Kauhl. 2013. Using Data Mining in Forecasting Problems. *SAS Global Forum 2013*, *085-2013*.

[2] Steven Bird, Edward Loper, and Ewan Klein. NLTK Project. http://www.nltk.org/.

[3] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

[4] NAHB/Wells Fargo Housing Market Index (HMI) http://www.nahb.org/reference_list.aspx?sectionID=134.

[5] China Industrial Production Index http://www.stats.gov.cn/english/statisticaldata../MonthlyData/.

[6] Consenses Economics GDP Forecast
http://www.consensuseconomics.com/Forecast_Surveys/Economic_Forecast_Probabilities.htm

[7] Lee, T., ET. All, (2008) "Two-Stage Variable Clustering for Large Data Sets," SAS Institute Inc., Cary, NC, SAS Global Forum, Paper 320-2008.

[8] Engle, R. and Granger W. (1992) Long-Run Economic Relationships: Readings in Cointegration, Oxford University Press.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Michael P. Dessauer
Organization: The Dow Chemical Company
Address: 2040 Dow Center
City, State ZIP: Midland, MI 48640
Email: mpdessauer@dow.com
Web: http://www.dow.com/dowservices/consulting/partners/advancedanalytics.htm

Name: Justin Kauhl
Organization: Tata Consultancy Services
Address: 2511 E Patrick Rd
City, State ZIP: Midland, MI 48640
Email: justin.kauhl@tcs.com
Web: http://www.tcs.com/offerings/business_process_outsourcing_BPO/analytics-insights/Pages/default.aspx