USING SOCIAL MEDIA AND ONLINE CONVERSATIONS TO ADD DEPTH TO UNEMPLOYMENT STATISTICS

Methodological White Paper[1]

December 8, 2011

Abstract

This research investigates whether and how social media and other online user-generated content could enrich understanding of the effect of changing employment conditions. The primary goal is to compare the qualitative information offered by social media with unemployment figures. To this end, we first selected online job-related conversations from blogs, forums and news from the United States and Ireland. For all documents, a quantitative *mood score* based on the tone of the conversations—for example happiness, depression or anxiety— was assigned. The number of unemployment-related documents that also dealt with other topics such as housing and transportation was also quantified, in order to gain insight into populations' coping mechanisms. This data was analyzed in two primary ways. First, the quantified mood scoring was correlated to the unemployment rate to discover leading indicators that forecast rises and falls in the unemployment rate. For example, the volume of conversations in Ireland categorized as showing a confused mood correlated with the unemployment rate with a lead-time of three (3) months. Second, the volume of documents related to coping mechanisms also showed a significant relationship with the unemployment rate, which may give insight into the reactions that can be expected from a population dealing with unemployment. For example, the conversations in the US around the loss of housing increased two (2) months after unemployment spikes. Overall, in this initial research, SAS and Global Pulse have underlined the potential of online conversations to complement official statistics, by providing a qualitative picture demonstrating how people are feeling and coping with respect to their employment status.

1. Background and Objective

Global Pulse is dedicated to better understanding how new types of data can strengthen available information on how people are impacted by global crises. This project, conducted in partnership with SAS, seeks to lay a foundation to use what Global Pulse believes could represent a powerful source of new data: the global conversation that is taking place over social media and online content.

---

[1] This methods white paper arose from an on-going series of collaborative research projects conducted by the United Nations Global Pulse in 2011. Global Pulse is an innovation initiative of the Executive Office of the UN Secretary-General, which seeks to harness the opportunities in digital data to strengthen evidence-based decision-making. This research was designed to better understand where digital data can add value to existing policy analysis, and to contribute to future applications of digital data to global development. This project was conducted in collaboration with SAS. For more information on this project or the other projects in this series, please visit: http://www.unglobalpulse.org/research.

In particular, this research focuses on what unemployment looks like in social media, specifically from June 2009 to June 2011.

This project investigates whether and how the feelings and information shared on social media and other, online, user-generated content could enrich understanding of the effect of changing employment conditions on people's perceptions and decisions.

The research zoomed in on the cases of the US and Ireland, where social media and blogs are widely used and where the recent and ongoing economic crisis has had a severe impact on employment. In particular, the robust economy and low unemployment in both the US and Ireland has, over the past two to three years, given way to unprecedented high unemployment rates (see Figure 1). In Ireland, according to EuroStat, the unemployment rate has been under 5% since 2000; since the economic crisis, the rate has risen to as high as 15% and fluctuated around 14%, a level not seen since the early 90s. According to the US Bureau of Labor Statistics, the unemployment rate in the US reached double digits for the first time since the early 80s, and has dropped slightly from that peak to hover around 9%. The organization responsible for providing the Ireland unemployment rates to the database is the Central Statistics Office, Dublin and for the United States, the data was provided by the United States Census Bureau.

In this exploratory project, two specific questions were addressed: can online conversations provide an early indicator of impending job losses, and can they help policy makers enrich their understanding of the type and sequence of coping strategies employed by individuals?
In order to answer these questions, job-related user-generated unstructured text data was analyzed using SAS tools and technologies (see **Annex A** for a detailed list). This allowed researchers to first retrieve pertinent documents; second, analyze those documents to quantify the feelings, moods, concerns and strategies; and finally to correlate the data with official unemployment statistics.

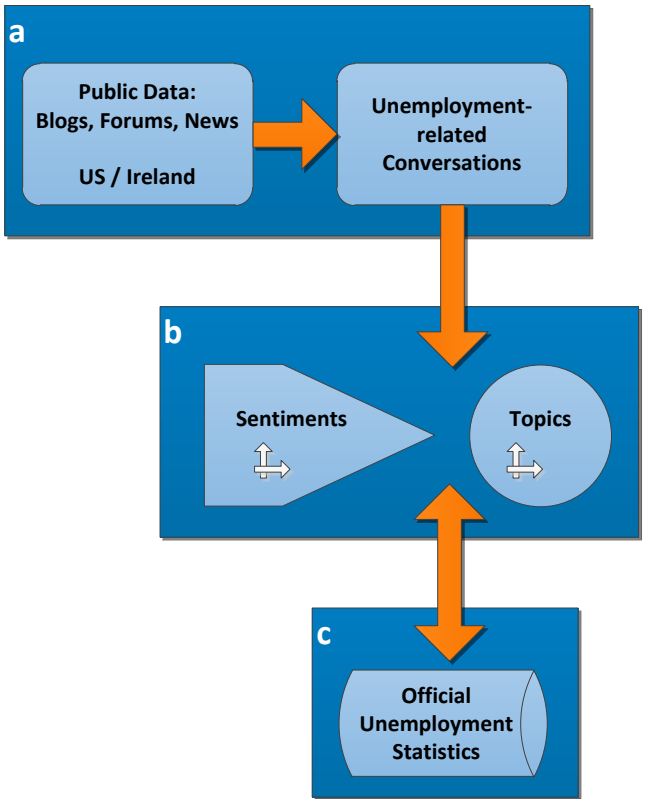

Figure 1: Project Workflow

2. Process Overview

The project was implemented according to following workflow (see Figure 1):

a) Online job-related conversations from blogs, forums and news were automatically retrieved.

b) Each document was assigned a quantitative *mood score* based on the tone or mood of the conversations—for example happiness, depression, anxiety—it contained. The number of unemployment related documents that also dealt with other topics, such as housing and transportation, was quantified and categorized into pre-defined lists of document topics representing potential coping mechanisms.

c) These measures—aggregated mood scores and the volume of conversations around different topics— were compared with official unemployment statistics over time in search of interesting dynamic-correlations.

3. Data Acquisition and Treatment

3.1. Acquisition

This process defines the relevant data set. The sample data was pulled from a variety of public Internet sources, i.e. blogs, internet forums and news, published in the US and Ireland over the past two years. Relevant documents containing references to terms defining unemployment,

specifically terms and phrases such as "unemployed, fired, on the dole, and collecting unemployment", were automatically retrieved and identified. The selection was subsequently refined using additional unemployment synonyms.

3.2. Filtering

Once the data set is defined, the documents pass through a custom filtering process meant to eliminate noise and sort the documents by topic.

The words used in each document were mined in order to assign one or more topical categories. These categories were chosen to represent some typical coping mechanisms and included: Housing, Transportation, Entertainment, Consumer Spending, Bills, Financial Distress, Underemployment, Nutrition Quality, Alcohol, Borrow/Save Money, Education, Healthcare, Unemployment Claims, and Travel (see **Annex B** for textual examples of these categories).

Categories related specifically to job status were also constructed, such as the chatter of people who *downgraded* by considering employment that paid less or required lower qualifications than previous employment. Finally, *underemployment*, where people are working only part time, was also examined.

3.3. Validation

Selected documents underwent a series of validation rounds to ensure the quality and accuracy of the query and filtering processes. For this purpose, a representative sample of the data was manually reviewed to check the relevance of the content and the adequacy of the categorization into different topical categories. Once a reasonable number of documents were successfully reviewed, the query and filtering processes were approved and the resulting dataset was deemed appropriate for analysis.

4. Analysis

4.1. Sentiment analysis

Each selected document was then evaluated with sentiment analysis techniques—a form of automated text analysis that assesses the nature (e.g. anxiety, confusion, hostility, etc.) and polarity (positive, negative, neutral) of the content expressed using natural language processing and linguistic rules. Each document was subsequently assigned a mood score, according to six different sentiments using the SAS' sentiment analysis engine (see Annex B2 for textual examples of these categories). When a word or phrase within a document is a predefined classifier for one of the mood categories, that document received a numeric score for that mood.

The application displays the various moods that dominated social media in any given month (as depicted on the left) or the change in a mood state over time (*confidence* in the US depicted on the right).

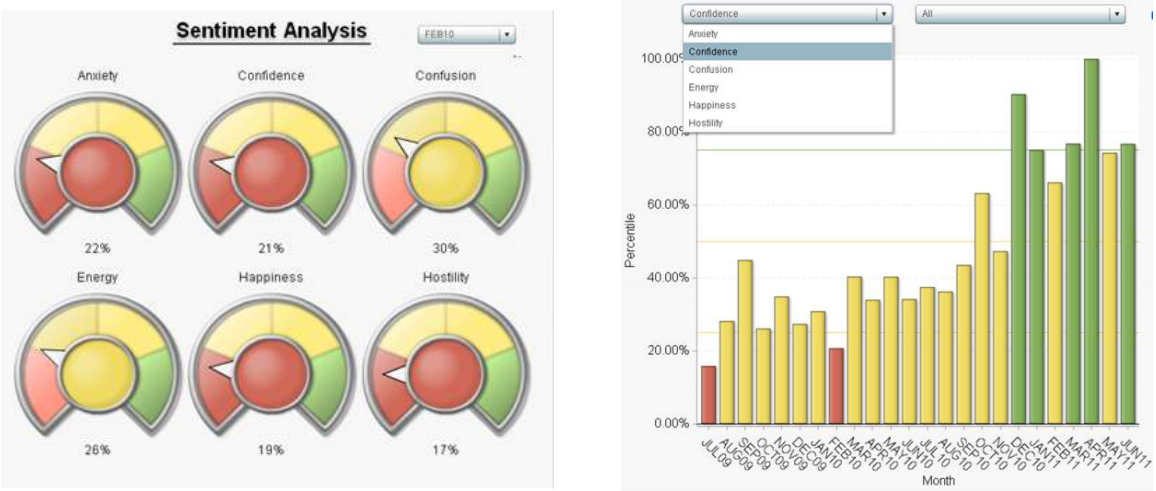

Figure 2: Sentiment Analysis

This analysis gives the user an idea of how the unemployed feel about their status. Are they optimistic about the future? Are they depressed or anxious about their job prospects? In this manner, sentiment analysis can provide an important complement to the unemployment statistics by essentially quantifying the qualitative experience of the unemployed.

4.2. Cross-Correlation Analysis

Finally, dynamic-correlation tests between mood scores and the volume of documents in a given category with official unemployment rate were performed in order to reveal, in social media and online user-generated content:

- Leading indicators that could give advance warning about unemployment statistics
- Lagging indicators of impact and coping strategies that could be used to design policy interventions aimed at mitigating the effects of unemployment, i.e. to assess how various types of unemployment chatter might predict the impact of changing employment conditions on other public services, such as the use of public transportation, housing, etc.

These dynamic-correlation tests were conducted using the maximum value of the average weekly moods scores assigned to the corpus of selected documents. Coping strategy volumes are obtained per month. These scores and volumes are tested for dynamic-correlation against the official unemployment rate. Only results at a 90% confidence level or higher are discussed here and presented in the dashboard.

The category of transportation provides one compelling example. The dynamic-correlation analysis plot below shows a statistically significant relationship between transportation chatter and the actual unemployment rate.

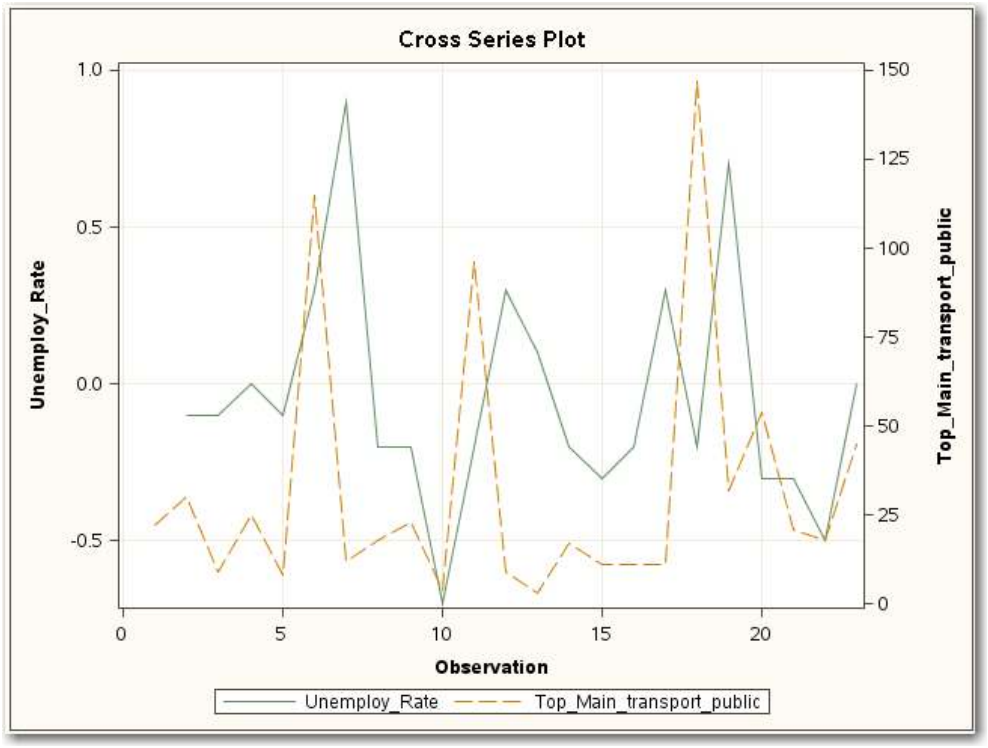

Figure 3: Cross Series Plot

Upon further investigation, it became clear that discussions about public transportation spike one month prior to a spike in unemployment. This suggests that an increased demand for public transportation can be expected to spike prior to a spike in unemployment.

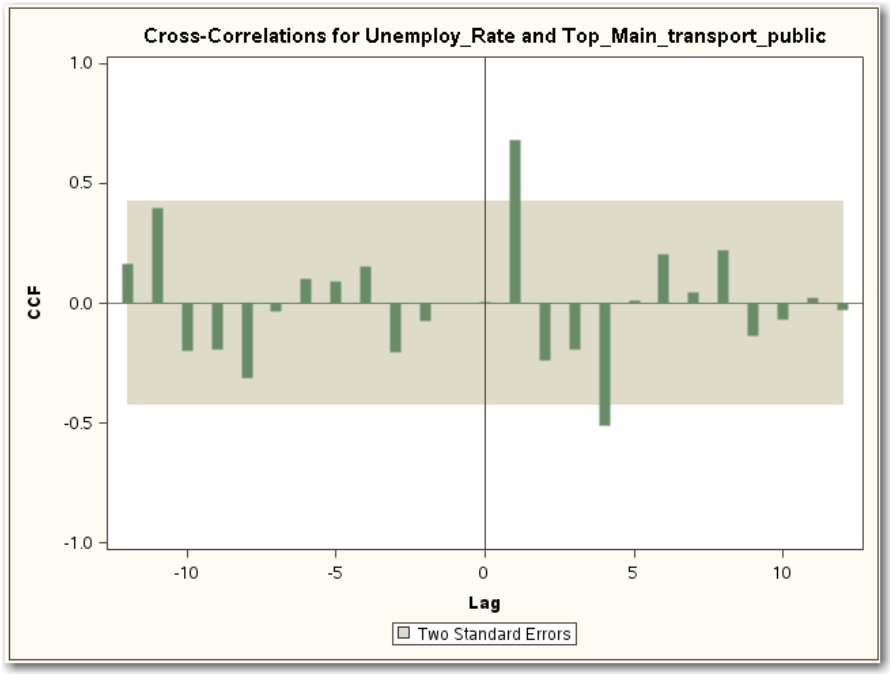

Figure 4: Cross-Correlations

Correlation analysis was conducted for all of the mood states and coping strategies and resulted in statistically significant correlations between several categories and the official unemployment rate (see Annex C for more examples of cross-correlation functions).

5. Outcomes

5.1. Unemployment Dashboard

The SAS team created two Dashboards—one for Ireland and one for the US— that provide a visual representation of findings. The dashboards were designed to be easily understood by non-experts, and it is an example of how one might monitor the real time breakdown of social listening results in the future.

The Dashboard allows the user to investigate the volume of conversations around unemployment as well as the coping mechanisms that are being discussed in relation to unemployment. The Dashboard also identifies time relationships between unemployment, conversation moods, coping mechanisms and various macroeconomic indicators, which allow the user to see patterns and predictions. Understanding the timing and circumstances that are causing a shift in a populations' behavior allows decision makers to prepare and act.

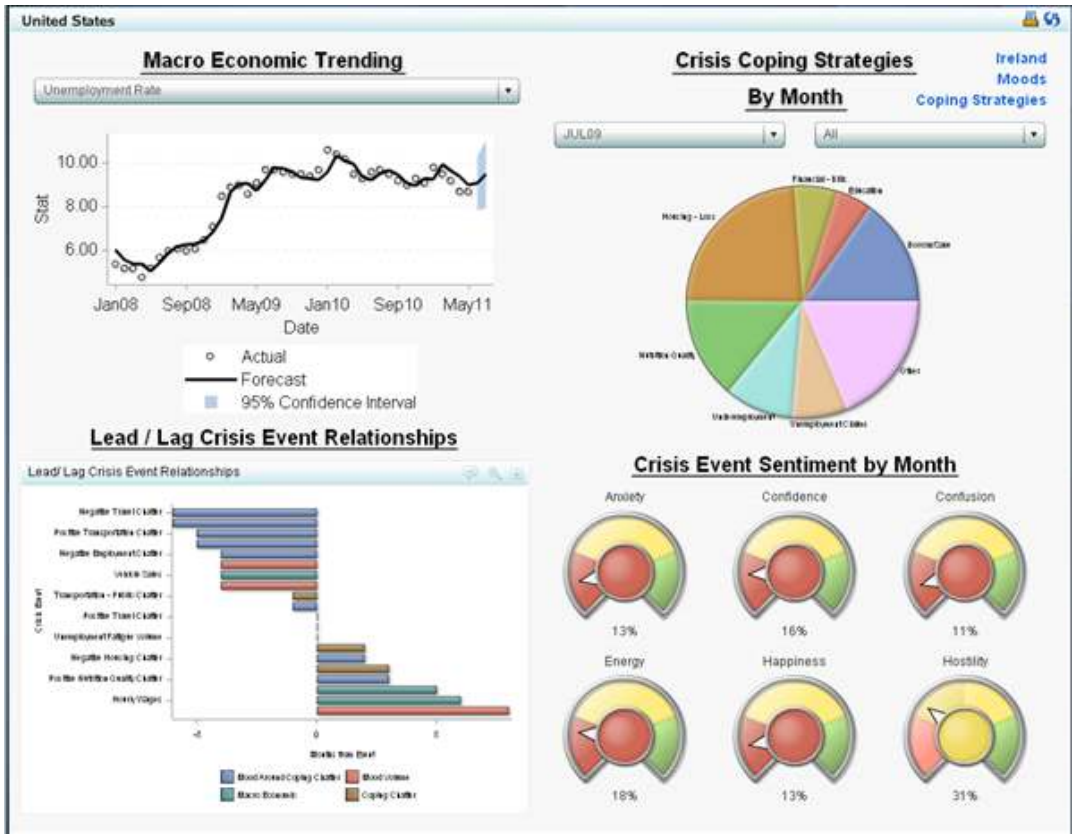

Figure 5: United States Dashboard

In the top left quadrant

The official unemployment rate and the forecasting model are depicted. The forecast of the unemployment rate was obtained using the SAS® Time Series Forecasting System. The model

with the best Akaike information criterion to measure the relative goodness of fit of a statistical model is chosen. A Linear Holt Exponential Smoothing model was chosen for the United States and a Winters Method Additive Model was chosen for Ireland. In the future, the results such as those from this proof of value study could become valuable inputs for a predictive unemployment model. This quadrant exemplifies the possible display of such a model.

In the top right quadrant

The volume of documents that discuss particular topics are presented per month in a pie chart.

In the bottom right quadrant

The mood scores per month are shown as six mood dials or barometers. The benefit of presenting the mood in this manner is that it allows a quick determination of changes over the past month—in other words; it is easy to see if there has been a negative or positive change in mood over the past month. This information is also available over time. By clicking on the title above the mood dials, a new dashboard will load which displays the results for a selected mood in a bar chart format over time.

In the bottom left quadrant

Dynamic-Correlation results are shown in a vertical bar chart. The time lag of the correlation is on the x-axis and the variable correlated with the unemployment rate is on the y-axis.

Other features of the dashboard allow the user to understand the analysis better. Additional links contained in the dashboard will:

- provide examples of phrases from actual document which describe the coping strategy or mood state
- allow the user to explore the dynamic-correlations in more depth

These features allow the user to understand better the nature of the documents and the data that informs the analysis.

Finally, the dashboard allows users to look at changes over time. Selecting a month changes all other appropriate visuals, allowing for the exploration of the interplay between the unemployment rate, the mood of the unemployed and coping mechanisms. The data and results are represented in a SAS® Business Intelligence Dashboard Interface.

5.2. Results

Starting from June 2009 to June 2011, the dashboard retrieved over 28,000 documents for Ireland and 430,000 for the United States. Over half of the documents pulled in each contained coping mechanism classifiers.

The following figures provide a graphical example of the leading and lagging relationships found in the data:

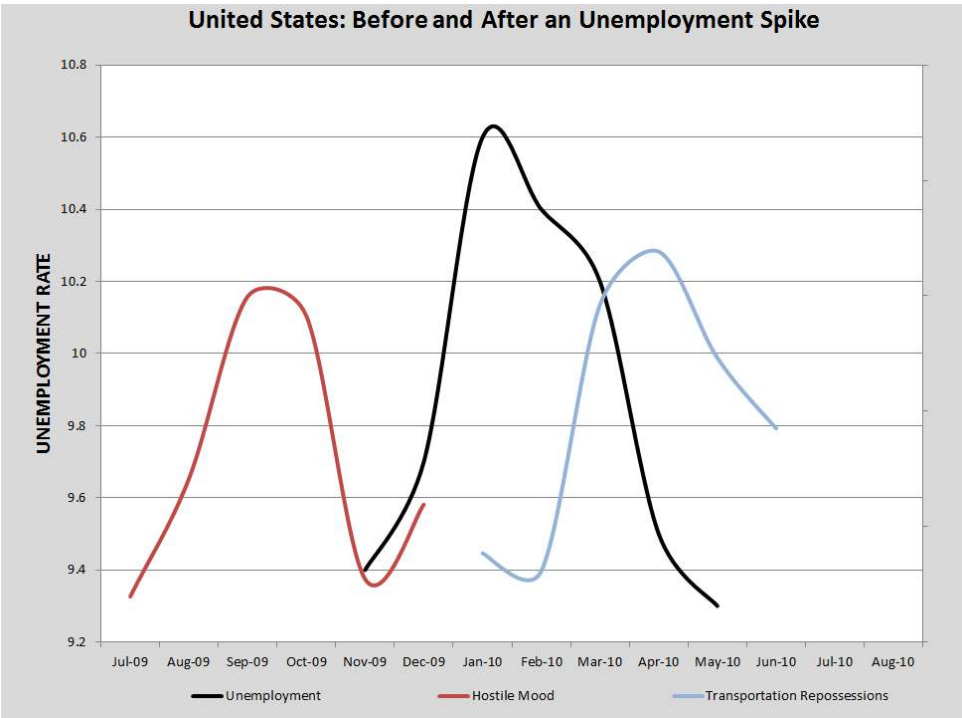

Figure 6: United States Unemployment Spike

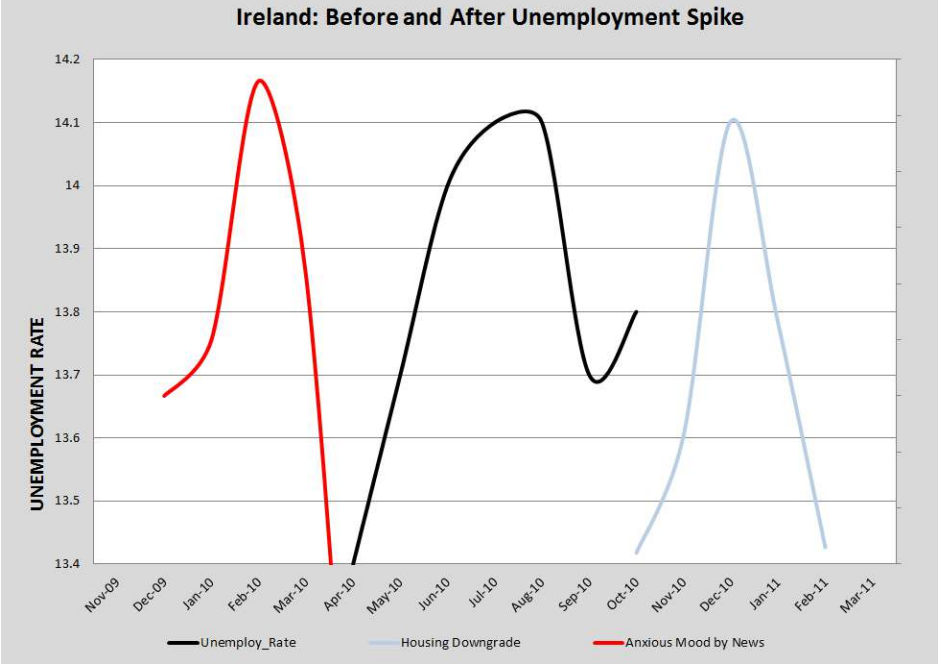

Figure 7: Ireland Unemployment Spike

Overall, among the 40+ cross-correlations explored, five (5) indicators in the US and six (6) indicators in Ireland showed a correlation significant at the 90% confidence level or higher.

These findings are a first step in this kind of research and represent a high potential source of complementary information to the unemployment statistics. The following images depict the strongest dynamic-correlations that were found:

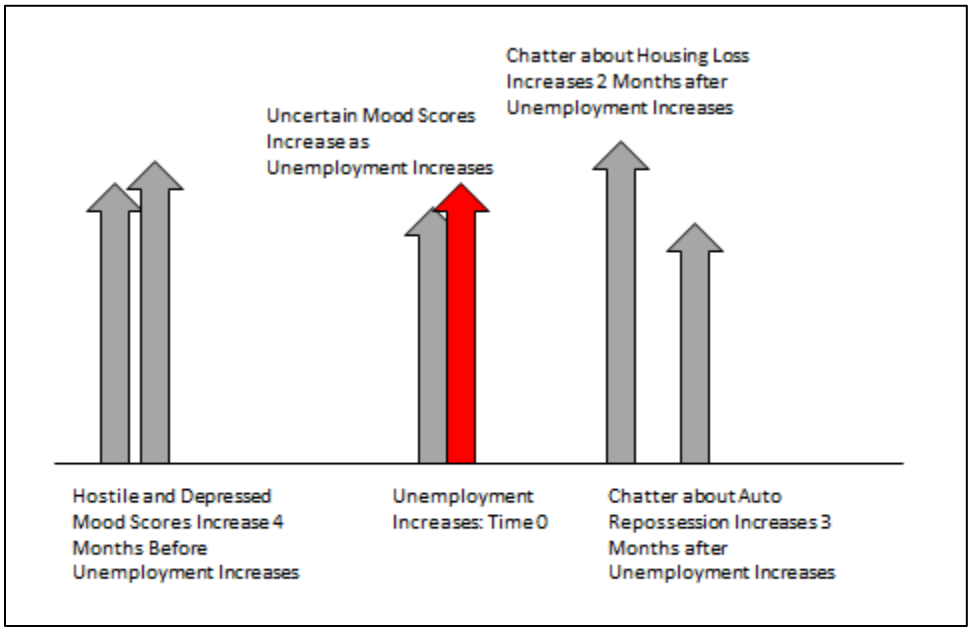

Figure 8: United States Chatter

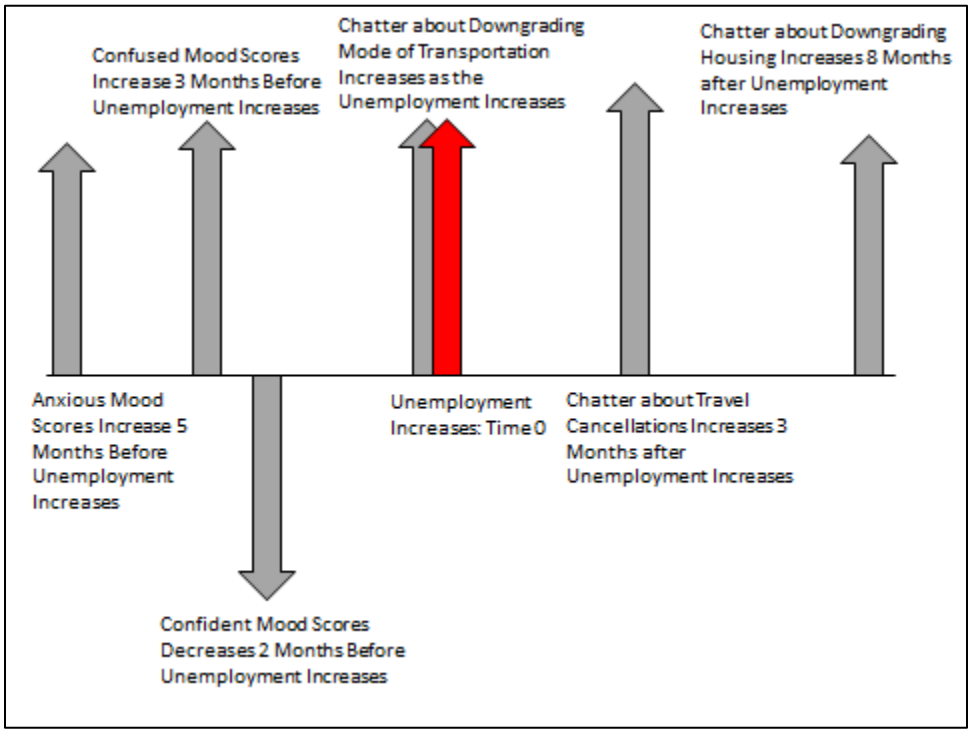

Figure 9: Ireland Chatter

Table 1: Country Correlation, CCF and Significance Level

Country	Correlation Description	CCF	Significance Level
US	Hostile Mood increases 4 months before a spike in unemployment	.442	95%
US	Depressed Mood increases 4 months before a spike in unemployment	.489	95%
US	Uncertain Mood increases as unemployment spikes	.448	95%
US	Talk about loss of housing increases 2 months after an unemployment spike	.634	95%
US	Talk about auto repossession increases 3 months after an unemployment spike	.455	95%
IRELAND	Anxious Mood increases 5 months before a spike in unemployment	.387	90%
IRELAND	Confused Mood increases 3 months before a spike in unemployment	.675	95%
IRELAND	Confident Mood decreases 2 months before a spike in unemployment	-.407	90%
IRELAND	Talk about changing transportation methods for the worse increases as unemployment increases	.380	90%
IRELAND	Talk about travel cancelations increases 3 months after an unemployment spike	.450	95%
IRELAND	Talk about changing housing situations for the worse increases 8 months after unemployment increases	.328	90%

6. Conclusions and perspectives

In this initial research, the high potential of online conversations to complement official statistics has been shown, by providing leading and lagging indicators that show how people are feeling with respect to their employment status and which coping strategies (conversation topics) are employed over time.

Several mood score volumes produced strong correlations with the unemployment rate, which may be leading indicators that the unemployment will rise or fall. For example, conversations in Ireland showing a confused mood preceded the unemployment rate variations by 3 months. And data pulled from the social listening sources, representing daily conversations, proved to contain valuable information related to how the unemployed cope. In addition, changes in the volume of particular coping topics showed significant lagging relationships with the unemployment rate that may give insight into reactions that can be expected from a population dealing with unemployment – as demonstrated, for example, by the increase in conversations in the US around the loss of housing 2 months after an increase in unemployment.

In short, the type of data needed to perform this sort of analysis exists, and the initial research pointed to potential indicators that might be used for improving unemployment policies or social protection programs.

Several challenges in this type of analysis and future lines of research were also made clear through the course of this work. First, greater geographical information related to each document would allow for a finer grain analysis at a regional level. In addition, one of the main challenges of any social media analysis is to understand better the demographic characteristics of the sampled population. Finally, online conversations are a relatively new source of data, and it is thus required to start maintaining a database of these documents over time, so in the future, analysis will cover a longer period.

The potential is high, and in the future, models that are more robust can be built for both helping predict and anticipate the consequences of unemployment spikes, and real-time social listening could be used to monitor the experience of the unemployment in an automated manner, continuously.

Annex A: Tools

Data acquisition: The data acquisition is handled through a third party service, Boardreader.

Data filtering: SAS® Content Categorization software is used to create the filter the relevant documents. Data must satisfy the standards set forth in SAS Content Categorization to be passed to the next step in the process.

Dynamic-Correlation analysis: Several tools have been used for the analysis, including SAS® Content Categorization Studio, SAS® 9.2 statistical and forecasting tools, SAS® Enterprise Guide and SAS® Sentiment Studio Workbench. The text is parsed through a SAS procedure called DOCPARSE. The SAS procedure, PROC TIMESERIES, is used to obtain the cross-correlation function results.

Mood State: Mood State is a method by which SAS measures the overall mood and specific moods of a data corpus. Unlike sentiment analysis, which is a simple positive/negative/neutral decision, mood state analysis offers a more refined measure by which to judge social media. Documents are scored to provide mood scores for Anxiety, Confidence, Hostility, Confusion, Energy, and Happiness.

Unemployment dashboard: The data and results are represented using the SAS Business Intelligence Dashboard Interface.

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 50,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.

Annex B

Example of documents categorization based on conversation topics

Topic Examples (all occur within same document as mention of Unemployment or synonym)

- Alcohol
  - "My beer budget will obviously be cut."
- Borrow/Save money
  - "I hate to do it but will have to get a payday loan."
- Consumer goods
  - "Definitely won't be buying new clothes this season."
- Nutrition quality
  - "Not sure how to afford fresh produce these days."
- Education
  - "We can't afford the tuition so she'll have to withdraw."
- Underemployment
  - "It's only part time but at least it's money coming in."
- Entertainment
  - "The kids will make do without new video games or dvds."
- Financial
  - "A few more months and we'll have to seriously consider a bankruptcy."
- Bills
  - "Sorry water bill, this month I pay the electric, next month it's the student loans."
- Healthcare
  - "He has cobra available but I don't know we'll ever pay for it."
- Housing
  - Loss
    - "That's it. The bank is foreclosing next week."
  - Downgrade
    - "We'll save a bunch if we can move to a smaller place."
- Transportation
  - Loss
    - "They're trying to repo my wife's car."
  - Downgrade
    - "I sold the Audi and got a used Mazda."
  - Public
    - "I canceled the car insurance so I'll start taking the bus. It's cheaper."
- Travel

- "The Easter vacation we planned is officially canceled due to lack of funds."
- Unemployment claim
  - "I filed for benefits the same day. Ughh."

Example of documents categorization based on sentiment analysis

- Anxiety
  - "I'm nervous that I won't find another job like this one."
- Confidence
  - "She doubts that we can afford to have two cars."
- Hostility
  - "It's outrageous that we were only given two weeks severance!"
- Confusion
  - "I've felt so scattered since the layoffs."
- Energy
  - "My husband has been so sluggish for the past few months."
- Happiness
  - "The counselor said it was normal to experience a sense of mourning after losing a job."

Annex C

Cross-correlations functions between mood states and/or conversation topics against the official unemployment rate.

United States

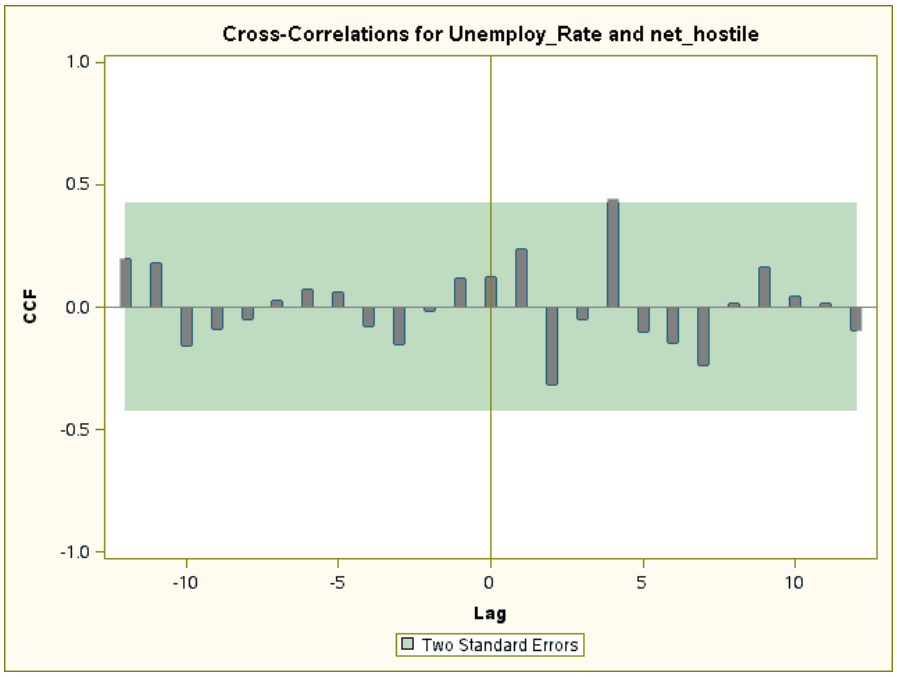

Figure 10: (Hostile Mood – Agreeable Mood) Hostility

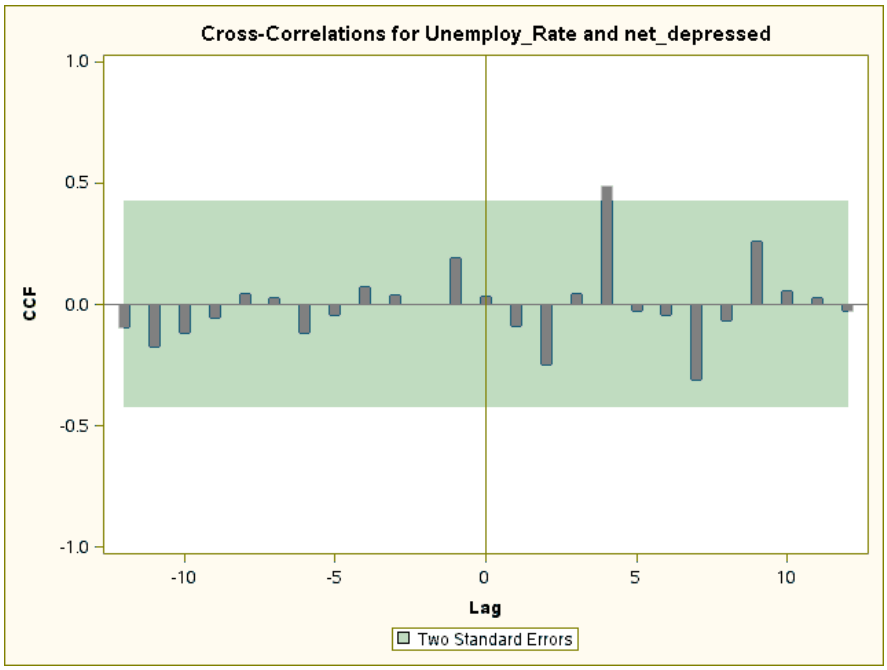

Figure 11: (Depressed Mood – Elated Mood) Depression

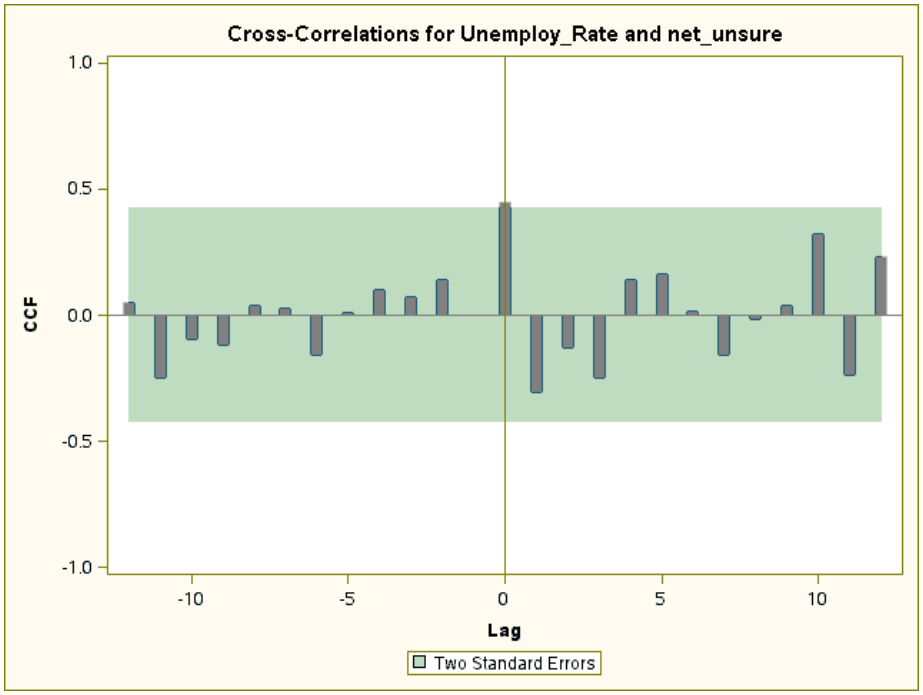

Figure 12: (Unsure Mood – Confident Mood) Uncertainty

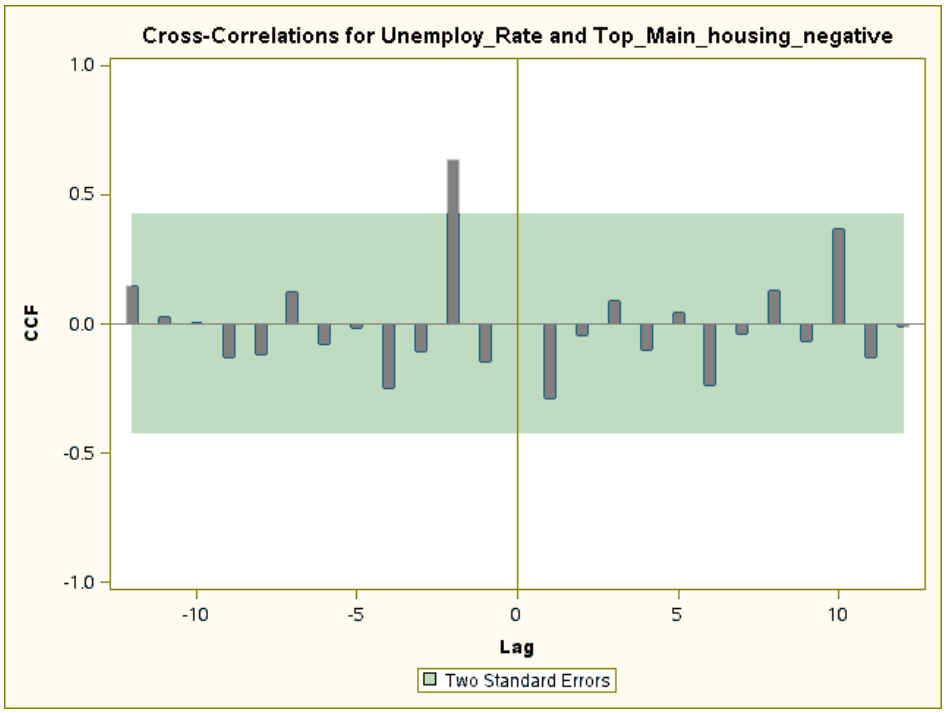

Figure 13: (Housing Loss)

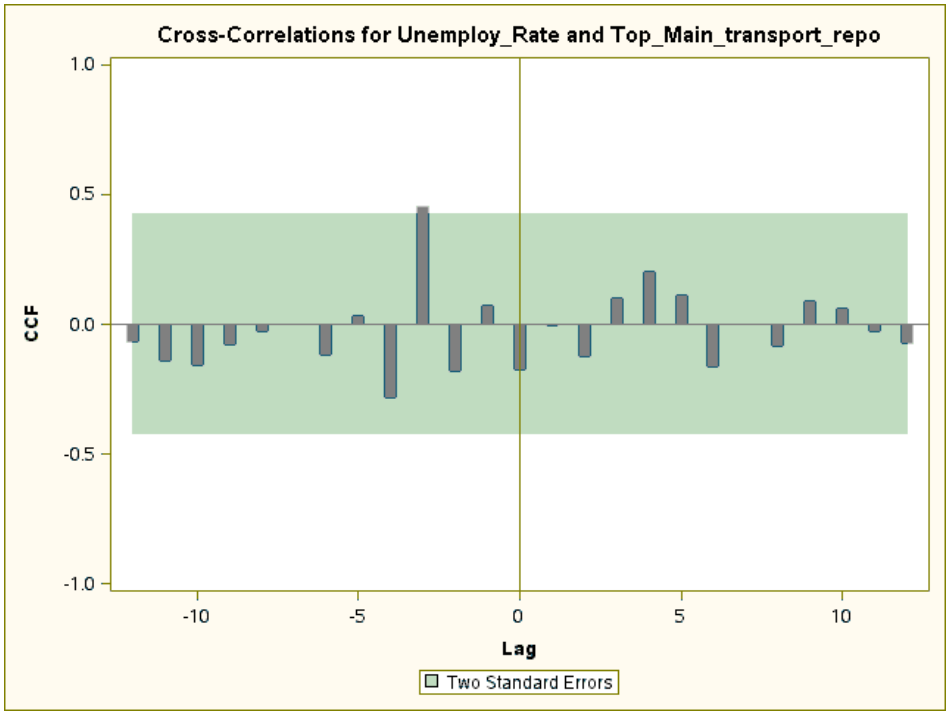

Figure 14: (Transportation Repossession)

Ireland

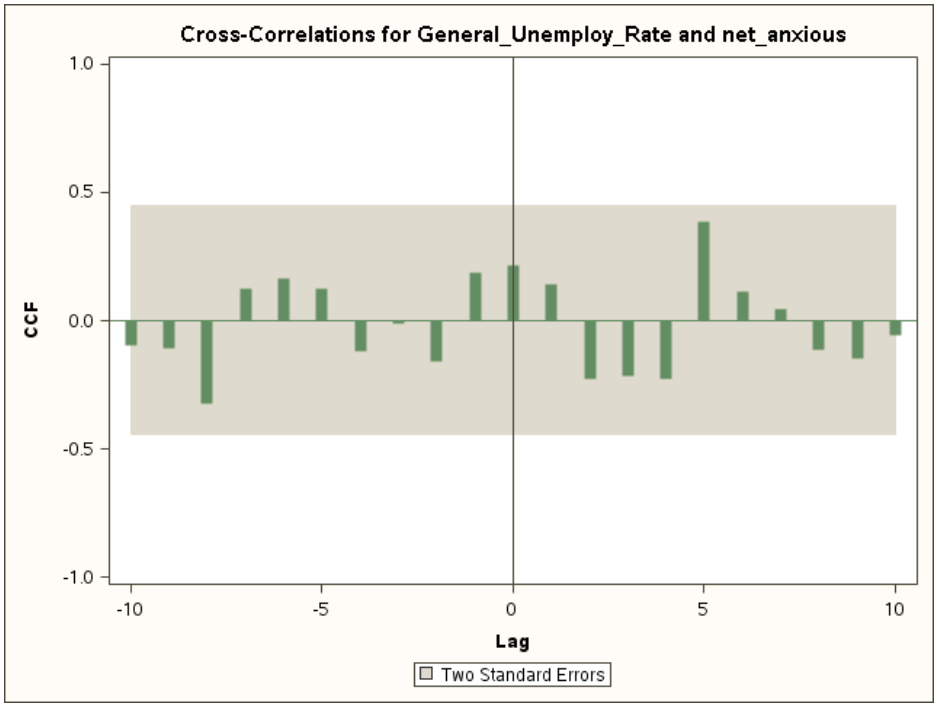

Figure 15: (Anxious – Composed News) Anxiety

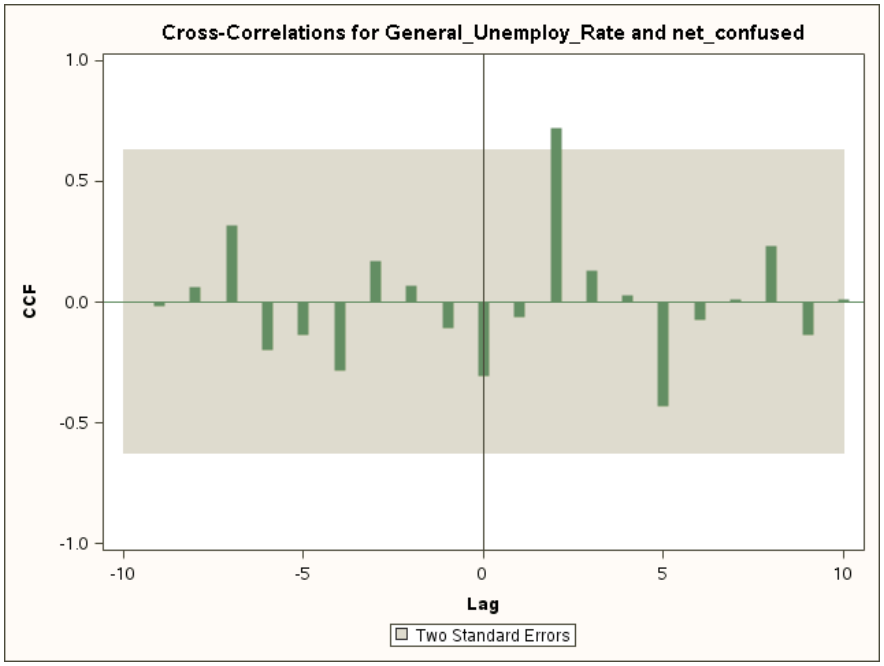

Figure 16: (Confused – Clearheaded) Confusion

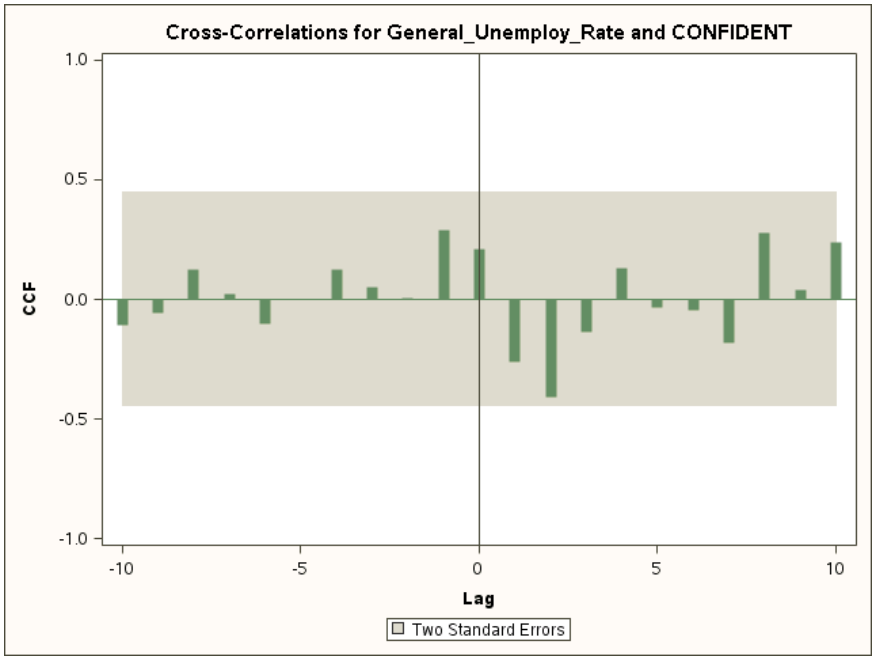

Figure 17: Confident by Blogs

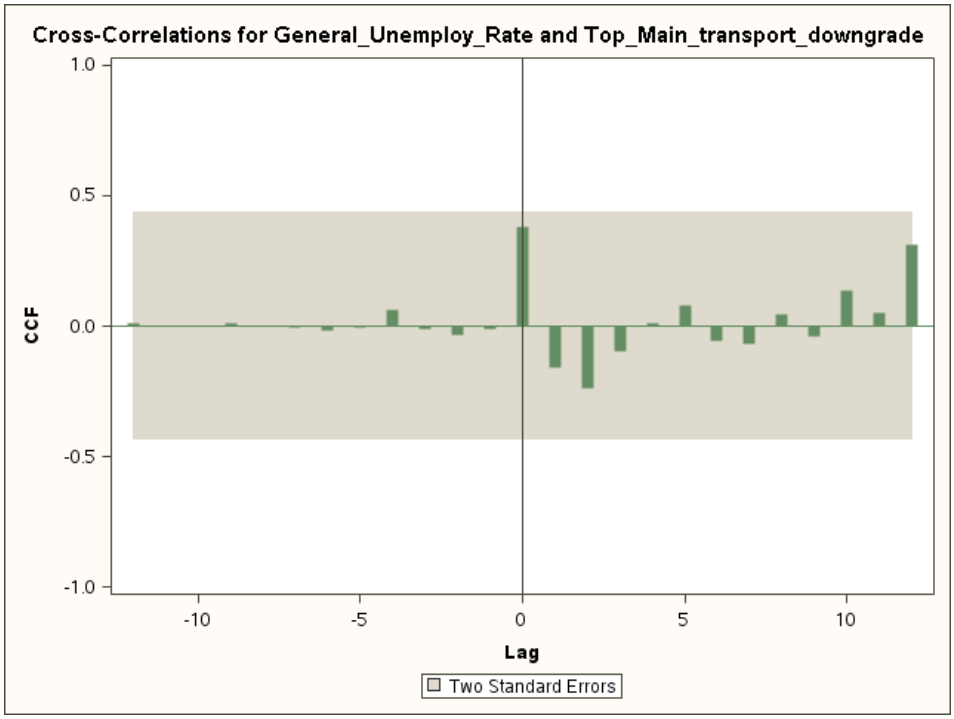

Figure 18: Transportation Downgrade

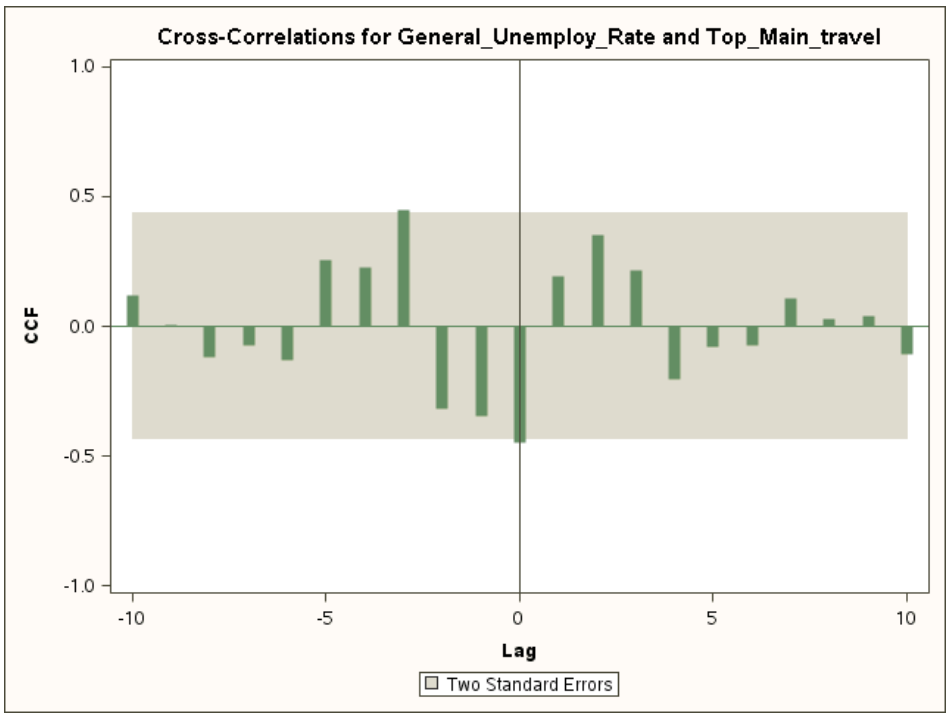

Figure 19: Travel Cancellations

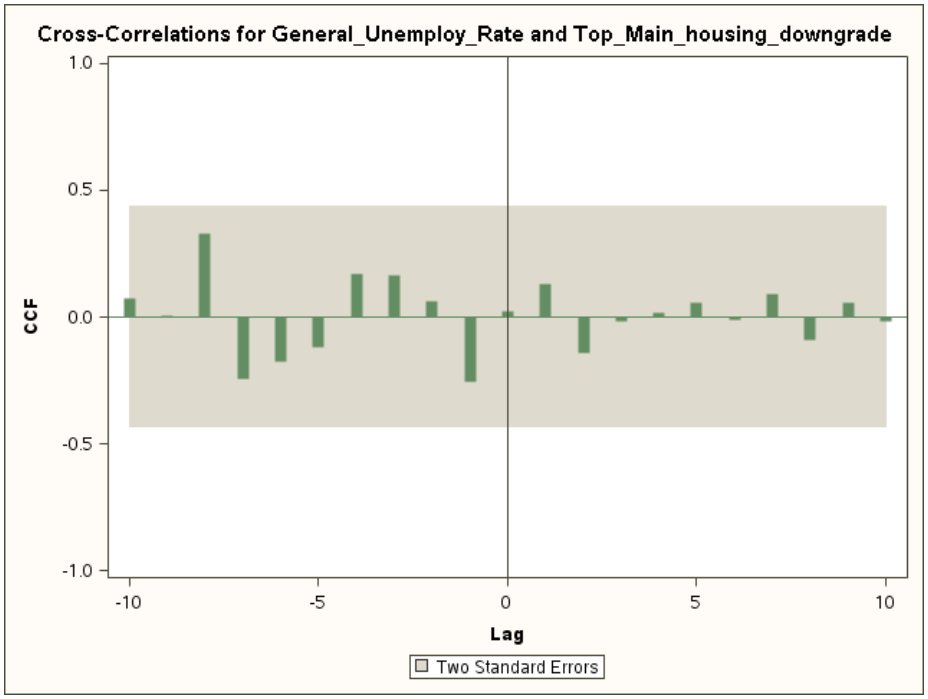

Figure 20: Housing Downgrade