IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

Probabilistic Methods for Enhancing Foreground Segmentation of Various Data Model using Big Data Model

Mahadevi Somnath Namose[1] and Dr. Tryambak Hiwarkar[2]
Research Scholar, Department of Computer Science[1]
Professor, Department of Computer Science[2]
Sardar Patel University, Balaghat, MP, India

Abstract: *Some indicators of social and economic health, especially those pertaining to developing countries, can swing wildly. A country's economy might take a hit if major economic indices like commodity prices, unemployment, currency exchange rates, etc., experience significant volatility. Instability in commodity prices is bad for economic development, financial reserves, and income distribution, and it may exacerbate poverty rather than alleviate it. Exports from various countries, including India, are dominated by commodities. The volatility of currency exchange rates has a ripple effect on commodity prices. Economic growth and stability require constant attention to these socioeconomic factors and an awareness of their inherent instability. Decades of research haven't shed any light on the reasons for a socioeconomic index's anticipated time and place fluctuations or the relationships between several indices. Economists can understand and foresee the volatility of social and economic indices with the use of predefined economic models. Traditionally, computational modelling has been the primary method of analysis for computer scientists when dealing with structured time series data. A rare opportunity to examine socioeconomic fluctuations has arisen because to the rapid expansion of unstructured data streams on the web and the development of cutting-edge computational linguistics algorithms during the past decade.*

Keywords: Probabilistic Methods

I. INTRODUCTION

Some indicators of social and economic health, especially those pertaining to developing countries, can swing wildly. A country's economy might take a hit if major economic indices like commodity prices, unemployment, currency exchange rates, etc., experience significant volatility. Instability in commodity prices is bad for economic development, financial reserves, and income distribution, and it may exacerbate poverty rather than alleviate it. Exports from various countries, including India, are dominated by commodities. The volatility of currency exchange rates has a ripple effect on commodity prices. Economic growth and stability require constant attention to these socioeconomic factors and an awareness of their inherent instability. Decades of research haven't shed any light on the reasons for a socioeconomic index's anticipated time and place fluctuations or the relationships between several indices. Economists can understand and foresee the volatility of social and economic indices with the use of predefined economic models. Traditionally, computational modelling has been the primary method of analysis for computer scientists when dealing with structured time series data. A rare opportunity to examine socioeconomic fluctuations has arisen because to the rapid expansion of unstructured data streams on the web and the development of cutting-edge computational linguistics algorithms during the past decade.

In order to interpret, explain, and anticipate socioeconomic index volatility, this thesis proposes a set of breakthrough big data analytics algorithms that automatically infer events, knowledge graphs, and prediction models from unstructured news streams.

Most content on the web is still just scattered about in random text files. It may be difficult to know where to start looking for specific details in such a large volume of content. We hope to do this with this thesis by depicting this mountain of data in a way that is both clear and complete. We present alternative methods of describing events by mining these texts for pertinent information. As a lower-dimensional representation of web papers reporting on events

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

and incidents going place around the world, the idea of events provides a far more precise form of information. Here we have isolated instances that can be analysed independently.

Four sub-problems must be solved before the ultimate goal can be reached:
(1) finding a succinct way to represent events from news articles;
(2) determining how events are related to one another;
(3) determining which proceedings are related with observed variations in an indicator of interest; and
(4) determining how this data can be joined to project prognostic replicas to forecast the unpredictability of a socio-economic indicator.

II. LITERATURE SURVEY

1. Gillenwater et al. [1]using threads as a means of locating inter-document connections. Documents can be connected to one another through logical chains that are referred to as threads. In order to identify the stages of development that are typically present in event sequences,
2. Allan et al[2] research, which consisted of picking out a single sentence from each occurrence inside a news topic and then scoring those lines before going on to the next item. Identifying and analysing the developmental patterns of the themes present in the text streams
3. Yan et al. [3]introduced an innovative new approach (ETS). ETS tries to offer a connected collection of summaries for each date along the timeline given a large corpus of time-stamped online articles that are related to a broad news query. Particular focus is placed on relevance, coverage, coherence, and cross-date variability in these summaries. ETS is significant because it enables you to read the news much more quickly while also improving your comprehension of what you read. This hypothesis is responsible for a major improvement that can be credited to the finding of new knowledge in the media, and more particularly, the identification of the causal linkages that exist between different events. Knowledge discovery works that are comparable in scale have also garnered a significant amount of attention. The construction of a knowledge base (KBC) based on hundreds of millions of web pages
4. Shahaf [4],work good for completing tales, but they tell us nothing about the events that take place before and after the journey that the protagonist takes in the story.
5. A. Schick et al[5] Over the past few years, one of the most popular study topics has been autonomous information extraction. A major advance has been made when it has become possible to recognise and track patterns of discussion that appear in multiple online publications, such as news reports. Models such as LDA and its many offshoots, such as Dynamic LDA, are among the most well-known and widely used in this area.

III. CLASSIFICATION PARTICULARS WEB CONTENT

Our plan's overarching objective is to make use of the wealth of information that is now accessible on the internet in order to mechanically construct a location-specific information portal. This will be accomplished by utilizing the information. This website will be able to provide a concise summary of the most significant climatological and agricultural patterns in any given location. The following design steps served as the basis for the creation of the fundamental components of our system.

3.1 Obtaining Information via the World Wide Web

Creating a list of questions to ask was helpful in sorting through the vast amount of papers that can be found on the internet to find the information that was most pertinent to the situation. In the subsequent stage of the procedure, some preliminary processing was carried out in order to validate the various sources of information**.**

- **Ignoble Text Set:** In order to successfully complete this assignment, one of the most important steps that needed to be taken was to map out a route on the internet through which we could get the required information. The selection of certain key terms connected with the issue, which was agriculture and the factors that influence it, such as climate, was one of the most significant tasks that needed to be done in order to get such information. This was one of the most important things that needed to be done in order to get such

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

information. Agriculture is highly dependent on a wide range of distinct factors, some of which are described in the following paragraphs. By using a set of terms that has been carefully picked, it is feasible to utilize such terms as queries in order to search for and uncover meaningful and contextual pages on the internet. Therefore, the first thing that needed to be done for this assignment was to decide which categories were the most significant, and then the next thing that needed to be done for this assignment was to construct a list of questions that appropriately characterized each of those categories. We were successful in recognizing six key categories that each have a considerable influence on the amount of agricultural products that are produced. There are seven components, and they are as follows: the soil, the water, the climate, the agricultural practices, and the crops, as well as the pesticides and fertilizers. It was decided to assign a number of pertinent keywords to each category so that the explanation could be made more explicit. For example, the condition of the soil is an important consideration in agricultural productivity. The soil's type, fertility, and moisture levels, among other characteristics, are among the most important factors in determining how agricultural practices should be carried out. In addition to this, a category might also be associated with a noteworthy occurrence. For example, "soil erosion" is a phenomenon that might take place with soil. This term alludes to the phenomenon.

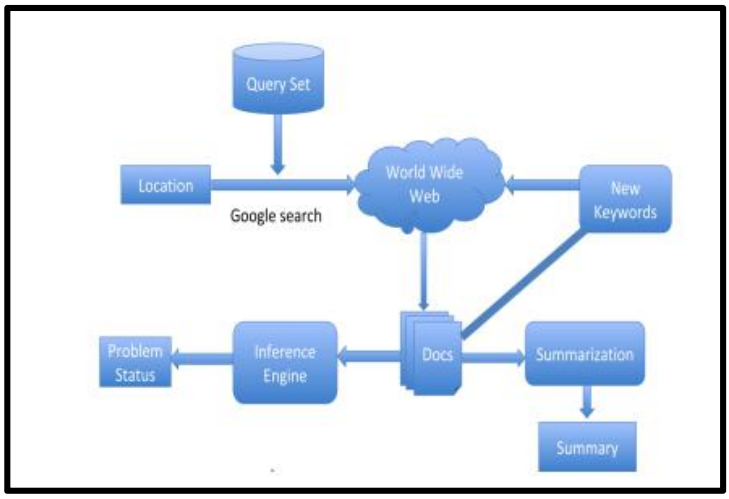

Figure 1: Architecture of the System

IV. INTERPRETATION ENGINE

From the language that was extracted from the documents, the purpose of the inference engine was to determine which issues were the most prevalent in a given location with regard to a particular category. In other words, provide answers to questions such as "Does Sambalpur suffer from water scarcity?" and "Are there any soil-related concerns in Jabalpur?"

The following set of heuristics were utilized in order to arrive at such crucial conclusions.

4.1 Presence of Keywords

A group of keywords is linked to each issue that is being discussed. Only a few of them were formatted by hand, while the remaining ones were automatically pulled from the source. The existence of such keywords in a piece of writing helps to validate the significance of the sense that is being conveyed by the query phrases. For instance, if the phrases "soil erosion" and "fluoride contamination" are used very frequently in a certain area, one can infer that there is a problem associated with either of those topics. Because we examined N-grams (N > 5) while extracting features, it is impossible to incorrectly conclude that there is a problem with soil erosion from the recurrent use of phrases such as "no soil erosion" or "soil erosion was not observed." In this scenario, the frequency of phrases such as "no soil erosion" or "soil erosion was not detected" would be higher than the frequency of the phrase "soil erosion."

The total number of occurrences obtained from a variety of sources

This metric offered a rough indication of the amount of people that were interested in the material. The diversity of the sources that reported the issue is one factor that can be used to infer the level of severity associated with the issue.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

Page Rank of the Pages

This statistic provides a quantitative assessment of the level of trustworthiness associated with the information source. If the credibility of the source can be established, it may be possible to eliminate the inclusion of trends that have been wrongly labelled as crucial.

V. METHODOLOGY

We take a look at the process that is followed in order to obtain a condensed answer to a web search query that was submitted by a user. Our summarization engine makes use of the Google Search API in advance to retrieve the top 64 (customizable) search result pages from Google. This allows us to get a head start. The primary objective of the summarization engine is to first do a thorough analysis of the text contained on each page of search results and then to aggregate all of that data into a single, condensed summary page. The final search response will be a condensed copy of the original text containing only the most crucial sections from the larger text, relevant images, and related connections to the many individual search result sites if the user feels forced to click on any particular result page. The summarization search engine is made up of three primary components: (a) the Text Summarization Engine, which condenses summaries across pages; (b) the Image Extraction Engine, which extracts relevant images from the page; and (c) the Aggregation and Presentation layer, which presents the condensed summaries as well as the extracted images. After that, we will delve even further into the specifics of each individual component.

5.1 Text Analyzer and Analyzer Engine

The act of reducing a lengthier piece of writing into a shorter version while maintaining the core concepts expressed in the original is referred to as the process of summarising. One can provide a summary of a piece of writing through a variety of strategies. Extraction-based summarizing and abstraction-based summary are two common methodologies that are utilized in the NLP literature. Summarization It is called "extraction-based summarizing," and it requires writing the summary by paraphrasing certain sentences from the original text. This is the approach that we take when we summarise materials. In order to accomplish this, it is necessary to select the most pertinent passages and phrases from the core text and make use of them when writing the summary. The conventional extraction-based summarising task is not the same as our own method of summarising the information. In contrast to the usual activity, the extraction procedure in this instance is founded on the terms of the inquiry. The objective of the summary form is to bring the reader's attention to the sections of the documents that have a considerable amount of relevance to the terms of the inquiry. Our extraction-based summarization method consists of two steps: (a) identifying key terms in a document in relation to the query terms; and (b) identifying key portions of a text by making use of the key terms that were discovered in the first step. Both of these steps can be found in the following paragraphs. It is possible to complete a work like this in the shortest amount of time and with the least amount of effort if you focus just on taking into account the sentences that are pertinent to the inquiry. On the other hand, taking such an aggressive approach to summarising can result in a significant loss in the task's level of accuracy. As a consequence of this, it is of the utmost importance to identify extra sentences within the text that contain significant information as early as is humanly possible. You will have an easier time understanding the parts of the text that contain a lot of information if you search for sentences in the text that contain these terms and then extract the information from those sentences in order to compose the summary.

5.2 Image Retrieval and Processing Engine

The process of extracting relevant photos connected to a query is not an easy operation in a web page since the visual organization of a website may have nothing in common with the real DOM structure of a page. As a result, the procedure might be complicated. In an ideal world, the image from a result page would only be extracted if it was pertinent to the search query and if its position on the page was close to the summary text that was provided by the NLP text summarizing engine for that page. However, this is not the case in the real world. We use a mix of two different parsing approaches in order to locate images that are relevant to a search query in the following ways: a) Techniques Relying on the Open Graph Protocol, and b) DOM Parsing.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

5.3 Assessment

A search engine that summarizes results aims to limit the amount of data that is transferred as well as the number of times that a user is required to enter search parameters in order to provide the user with the most relevant results to their query that can be found on the internet. It's possible that this would improve the user experience, given that any online contact in underdeveloped countries faces high latency owing to limited infrastructure. The summary search engine is designed to function in conjunction with the web search tool that you already use. It accomplishes this by utilizing the strategies that the underlying search engine use for ranking and indexing material as well as conducting searches. As a result, the objective of this evaluation is one of a kind in comparison to more conventional ways of information retrieval.

The effectiveness of a search engine that summarises results can be evaluated based on how quickly it provides those results and on whether or not it decreases the number of times a user must interact with the web in order to obtain the information they need.

In order to conduct an evaluation of our technology, we need to compile a comprehensive list of questions that faithfully depict its various applications.

5.4 Affinity Graph Label-Value Pair Discovery

If you look at the orientation of the words in the affinity network, you'll discover that the vertices might be descriptive, binary, or numeric. This is something you'll notice if you pay attention to the orientation of the words. Regarding the classification in question, we have a hypothesis that is founded on the following postulates:

- Descriptive keys will be represented on the graph as a component in the shape of a star; the central word will function as the key, and the phrases that surround it will function as the value for that key.
- A strong connection will exist between the nun vertex and the numerical keys. Because the values of the numeric keys are not consistent or standardised, this approach can only identify the keys for example, the price of a car is not a fixed entity, whereas for descriptive keys, like the colour of a car, have fixed values. Ads need to be processed in order to obtain their aggregate value for a certain numeric key.

The edge weights of each of the words that make up a binary-keyed component will be the same, and the component itself will come together to form a subgraph that has a high level of coherence.[9] A binary key cannot have several values associated with it at the same time. It is possible to deduce the values of binary digits based on their use as keys.

Algorithm 2

procedure GETKEYVALUE
Input: Affinity graph
Output: Set of keys and their values
    $Labels_{Desc} \leftarrow \{\}$; $Labels_{Bin} \leftarrow \{\}$; $Labels_{Num} \leftarrow \{\}$
    for each w_i in W **do**
        $score(w_i) = 0$
    end for
    for each w_i in W **do**
        for each x in $neighbor(w)$ **do**
            $P(x|w) = \frac{1}{deg(w)}$
            $score(x) = score(x) + P(x|w)$
        end for
        for each $e(w_i, w_j)$ in E **do**
            if $score(w_i) \gg score(w_j)$ **then**
                $Labels_{Desc}.add(w_i)$; $Value[w_i] \leftarrow w_j$
            else $score(w_i) \ll score(w_j)$
                $Labels_{Desc}.add(w_j)$;$Value[w_i] \leftarrow w_i$
            end if
            if $score(w_i) \approx score(w_j)$ **then**
                $Labels_{Bin}.add(w_i)$
                $Labels_{Bin}.add(w_i, w_j)$
            end if
            if $score(w_i) \approx score(w_{num})$ **then**
                $Labels_{Num}.add(w_1)$
            end if
        end for
    end for
    return $Label_{Desc}, Label_{Bin}, Label_{Num}$
end procedure

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

5.5 Performance

Using a corpus of 10,000 Craigslist auto advertisements and 8,000 apartment rental ads, a graph-based technique was used to learn a set of keys (descriptive, binary, and numeric) for the two distinct themes. The corpus was used to train the graph-based approach. The training keys were applied to a database that contained 2,984 advertisements for automobiles and 2,784 listings for apartment rentals so that the effectiveness of the training keys could be evaluated. The "golden set" was constructed by using key-value pairs that were painstakingly collected by hand from test sets. An experienced human annotator went through the training and testing data and assigned the descriptive, binary, and numerical labels to the golden set after going through the data. We were able to evaluate the unsupervised method by first extracting key-value pairs from the testing set using the affinity network that we had constructed during the training phase. Performance was evaluated utilising precision-recall values by contrasting the manually constructed golden set with the key-value pairs that were pulled from the database. [10] The F-values that were computed based on the precision and recall values for the automotive and apartment ad sets, respectively, are depicted in Figures 2 and 3, respectively.

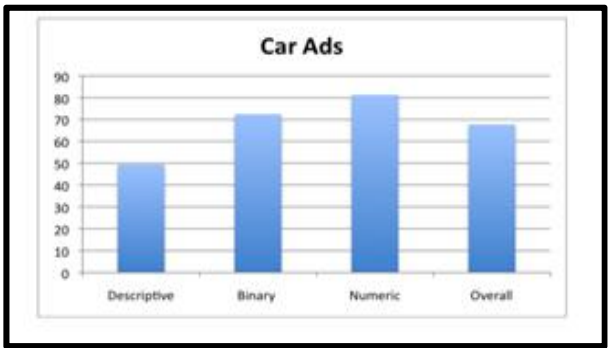

Figure 3: Accuracy for car ads

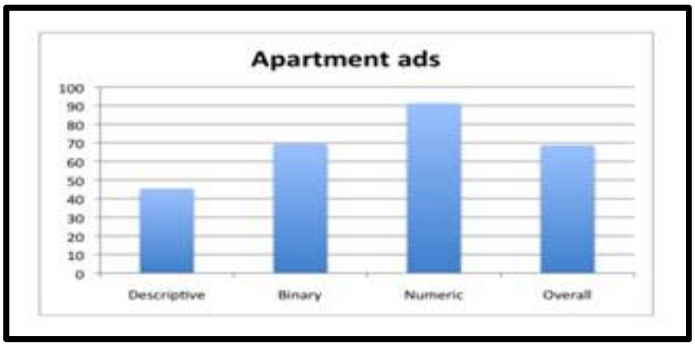

Figure 4: Accuracy for apartment

VI. PREDICTIONS OF FUTURE EVENTS

Multiple macroeconomic indices are estimated using data from a wide range of sources. However, in research trying to estimate such variables [7], analysis and forecasting are often carried out using structured data sources, taking into account just a small number of such parameters, which are also chosen manually.

The indices' extreme swings could be due to unknown reasons or a convergence of known and unknown ones that increase volatility. By monitoring events on a global scale, we can learn more about the sensitivity of these indicators. Reporting on such events in the real world is frequently done through unstructured text streams like the news, blogs, social media, etc.

In contrast to prior efforts, we (a) do not assume anything about the specific events that influence fluctuations of a given index and instead seek to automatically discover them; (b) do not use an external knowledge base or data to construct our model; and (c) restrict the scope of our prediction to a small set of variables. Given a large dataset of events related to the index of interest, we hope to be able to create a compact index-specific event-based predictive model to anticipate future changes in the index. One crucial premise of current attempts to use econometric models to explain the volatility of macroeconomic indices is that the model's underlying variables are known. Additionally, prior

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

algorithms for forecasting shifts in macroeconomic indices using news information have typically relied on pre-defined domain-specific features or market sentiment extraction procedures .

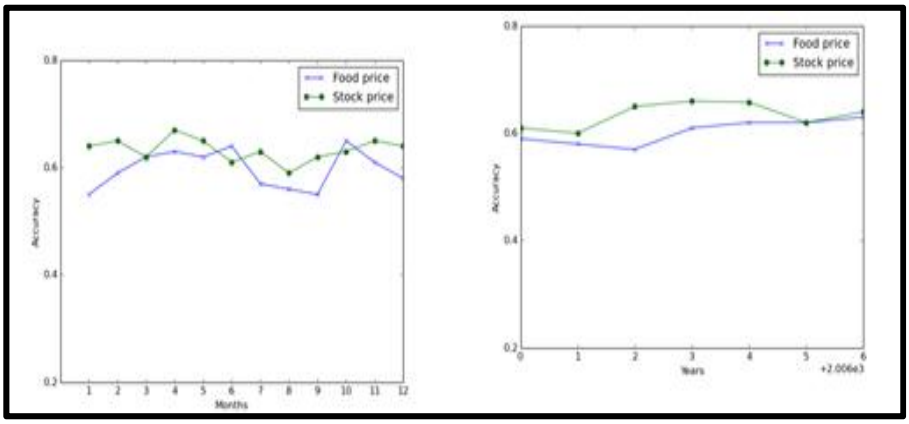

Figure 5: Monthly average Accuracy

VII. CONCLUSION

To that aim, we looked at the viability of using news articles as a source for event extraction and built knowledge graphs to represent event dependencies for usage in a wide range of analytics. We can make stronger and more credible claims if we have access to a wider range of news data and if we explore other research methods. This argument misses the mark in several crucial respects because it fails to adequately account for several relevant aspects of the data. One such bias could be that of a news organization. It's important to keep in mind that different media have different affiliations and interests, which can influence the news items they choose to cover and how they present those stories.

REFERENCES

[1]. J. Gillenwater, A. Kulesza, and B. Taskar. Discovering diverse and salient threads in document collections. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural LanguageProcessing and Computational Natural Language Learning, EMNLP-CoNLL ’12, pages 710–720, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[2]. J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’01, pages 10–18, New York, NY, USA, 2001. ACM.

[3]. R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. Zhang. Timeline generation through evolutionary trans-temporal summarization. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11, pages 433–443, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4]. D. Shahaf and C. Guestrin. Connecting the dots between news articles. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’10, pages 623–632, New York, NY, USA, 2010. ACM.

[5]. A. Schick, M. Bauml, and R. Stiefelhagen. Improving foreground segmentations with probabilistic superpixelmarkov random fields. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pages 27–31, June 2012.

[6]. Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD ’05, pages 198–207, New York, NY, USA, 2005. ACM

[7]. M. Michelson and C. A. Knoblock. Creating relational data from unstructured and ungrammatical data sources. J. Artif. Int. Res., 31(1):543–590, Mar. 2008

[8]. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. NIPS ’13, pages 3111–3119

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, January 2023
Impact Factor: 7.301

[9]. S. Nallareddy and M. Ogneva. Predicting restatements in macroeconomic indicators using accounting info, 2014

[10]. Y. Nishihara, K. Sato, and W. Sunayama. Event extraction and visualization for obtaining personal experiences from blogs. In Symposium on Human Interface and the Management of Information., pages 315–324, 2009.

[11]. D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. SIGIR '03, pages 235–242, 2003.

[12]. K. Radinsky and E. Horvitz. Mining the web to predict future events. In Proceedings of the sixth ACM international conference on Web search and data mining, pages 255–264. ACM, 2013.

[13]. D. Richards. Political complexity: Non linear models of politics. J. Artificial Societies and Social Simulation, 5(1), 2002.

[14]. A. Schick, M. Bauml, and R. Stiefelhagen. Improving foreground segmentations with probabilistic superpixelmarkov random fields. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pages 27–31, June 2012.

[15]. R. P. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Transactions on Information Systems (TOIS), 27(2):12, 2009.