Using Social Media Networks for Measuring Consumer Confidence: Problems, Issues and Prospects

By

Jane-Vivian Chinelo Ezinne Igboayaka

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for the degree of

Master of Science
in
Electronic Business Technologies

Faculty of Engineering
University of Ottawa
Ottawa, Ontario

Abstract

This research examines the confluence of consumers' use of social media to share information with the ever-present need for innovative research that yields insight into consumers' economic decisions.

Social media networks have become ubiquitous in the new millennium. These networks, including, among others: Facebook, Twitter, Blog, and Reddit, are brimming with conversations on an expansive array of topics between people, private and public organizations, governments and global institutions. Preliminary findings from initial research confirms the existence of online conversations and posts related to matters of personal finance and consumers' economic outlook.

Meanwhile, the Consumer Confidence Index (CCI) continues to make headline news. The issue of consumer confidence (or sentiment) in anticipating future economic activity generates significant interest from major players in the news media industry, who scrutinize its every detail and report its implications for key players in the economy. Though the CCI originated in the United States in 1946, variants of the survey are now used to track and measure consumer confidence in nations worldwide.

In light of the fact that the CCI is a quantified representation of consumer sentiments, it is possible that the level of confidence consumers have in the economy could be deduced by tracking the sentiments or opinions they express in social media posts. Systematic study of these posts could then be transformed into insights that could improve the accuracy of an index like the CCI. Herein lies the focus of the current research—to analyze the attributes of data from social media posts, in order to assess their capacity to generate insights that are novel and/or complementary to traditional CCI methods.

The link between data gained from social media and the survey-based CCI is perhaps not an obvious one. But our research will use a data extraction tool called *NetBase Insight Workbench* to mine data from the social media networks and then apply natural language processing to analyze the social media content. Also, *KH Coder* software will be used to perform a set of statistical analyses on samples of social media posts to examine the co-occurrence and clustering of words. The findings will be used to expose the strengths and weaknesses of the

data and to assess the validity and cohesion of the *NetBase* data extraction tool and its suitability for future research.

In conclusion, our research findings support the analysis of opinions expressed in social media posts as a complement to traditional survey-based CCI approaches. Our findings also identified a key weakness with regards to the degree of 'noisiness' of the data. Although this could be attributed to the 'modeling' error of the data mining tool, there is room for improvement in the area of association—of discerning the *context* and *intention* of posts in online conversations.

Acknowledgements

Above all, I want to thank God Almighty, for the strength, good health and patience He gave me through the course of this thesis write up.

I want to express my sincere gratitude to Professor Kindra, my supervisor for his continued support even though it has not been an easy journey.

I would like to thank my co-supervisor, Professor Mulvey who has motivated me all through writing this thesis. His patience, support, understanding, persuasion, knowledge, guidance and encouragement helped me push through even when I had given up. I want to say thank you.

I also want to acknowledge my husband, Ifeanyi who has given me enormous and utmost support—moral, spiritual and physical, love, guide, encouragement, motivation and continuous prayers in making sure I achieve my dreams and pursue my goals to reach the next phase of my life.

My sincere thanks go to my parents, Engr. (Sir.) Martin and (Lady) Evelyn Igboayaka for seeing me through my education and life. Their unrelenting prayers, love and support—in every way that gave me the inspiration to finish this research and thesis. Also, I want to thank my siblings; Sandra, Jennifer, Martin, Marcel and Sophia Igboayaka for their prayers and support.

Last but not the least, I want to thank all my friends and well-wishers; Chinenye, Edidiong, Susan, Chidi, Joshua, Theresa, Chinedu, Chrestina, Akhere, Chidinma, Dafe, Abdul, Uzoma, Oduwa, Ebere, Chinweike, Dorcas and Chigozie for their encouragement and support during the period of this thesis write up.

Table of Contents

Abstract ..... **ii**

Acknowledgements ..... **iv**

Table of Contents ..... **v**

Table of Figures ..... **vii**

List of Acronyms ..... **ix**

Chapter 1 Introduction **1**

1.1. An introduction to established methods of Consumer Confidence Index measurement 4

1.2. Illustration of the use of Consumer Confidence Indices 4

1.3. Social Media Networks as a source of data for Consumer Confidence measurement 6

1.4. Research Objective 6

1.5. Research Question 6

1.6. Research Motivation 7

1.7. Research Approach and Contribution 7

1.8. The structure of thesis 9

Chapter 2 Literature Review **10**

2.1. The economic indicator of consumer confidence index 10

2.3. Measurement of Consumer Confidence 18

2.4. Social media networks 19

2.5. Attributes of Social Media Networks 22

2.6. Data mining concepts 24

2.6.1. Natural Language Processing concept 26

2.6.2. The Linguistic Inquiry and Word Count concept 28

2.6.3. Life History Calendar (LHC) 28

2.6.4. Method Proposed by Nigam and Hurst 29

2.7. Data mining tool 30

2.7.1. About NetBase® 30

2.7.2. Advantages of Using NetBase® 31

2.7.3. Limitations of NetBase® 31

2.8. Chapter Summary 32

Chapter 3 Methodology **33**

3.1. Social media data collection 33

3.2. Keywords for search 34

3.3. Research data analysis method 36

3.3.1. KH Coder 37

3.3.2. Raw dataset classification criteria 39

Chapter 4 Results and Analysis **41**
4.1. Assessing the robustness of social media networks data 45
4.1.1. Part A: Qualitative analysis based on Infographics 45
4.1.2. Part B: Quantitative analysis of aggregated summaries 60
Tax: Top terms 61
4.1.3 Part C: An analysis of SMNs content contribution 78
4.1.4 Part D: Analysis of raw data indexed by NetBase 83
4.2. KH Coder: Exploration of natural attributes of raw data indexed by NetBase 87
Chapter 5 Summary & Conclusions **91**
5.1. Summary 91
5.2. Conclusion 93
5.3. Future Study 95
Bibliography **96**

Table of Figures

Fig. 1: A comparative illustration of the correlation between the CCI and other economic indicators 5
Fig. 2: Example of the architecture of a typical opinion mining application used for a product review (Tsytsarau and Palpanas 2012) 8
Fig. 3: Differences between the three oldest forms of consumer confidence index measurement (Source: Reserve Bank of Australia) 11
Fig. 4: A survey of restaurant customers on their assessment of their personal finances 14
Fig. 5: A global CCI of sorts generated from a survey of 60 countries for the period of October – December 2013; Source: Nielsen, Consumer Confidence Concerns and Spending Intentions around the World Quarter 4, 2013. 19
Fig. 6: Global Penetration of social platform-Account ownership and Active usage— (Chaffey, Dave, 2014) .. 24
Fig. 7: Illustrating the association of words with other words used with them, an example of co-occurrence network of words. Source: (Higuchi, 2013) 38
Fig. 8: Illustrating the association of words with correspondence analysis. Source: (Higuchi, 2013) 38
Fig. 9: An example of output from hierarchical analysis using KH Coder. Source: (Higuchi, 2013) 39
Fig. 10: Word cloud of (a) Top Emotions expressed; (b) Top Terms used and (c) Most used hash tags 41
Fig. 11: Qualitative summary attributes of data associated with Tax keyword 48
Fig. 12: Qualitative summary attributes of data associated with the Capital Expenditure keyword 52
Fig. 13: Qualitative summary attributes of data associated with the Employment Condition keyword 55
Fig. 14: Qualitative summary attributes of data associated with the Health keyword 56
Fig. 15: Qualitative summary attributes of data associated with the Consumer Confidence keyword 57
Fig. 16: Qualitative summary attributes of data associated with the Political Affiliation keyword 58
Fig. 17: Qualitative summary attributes of data associated with the Retirement keyword 59
Fig. 18: Graphed Top term representation of the quantified attributes of data from social media networks, associated with Tax 63
Fig. 19: Quantified representation of the top emotions expressed that are related to Tax 67
Fig. 20: The hashtags used in conversations that are contextually associated with Tax 69
Fig. 21: The volume of top terms associated with Capital Expenditure 73
Fig. 22: Graphical representation of the number of Top emotions expressed, relative to Capital Expenditure 76
Fig. 23: The most used hashtags related to the subject of Capital Expenditure 77
Fig. 24: The top social media networks in terms of the volume of conversations 82
Fig. 25: Co-occurrence associations of indexed raw datasets 88
Fig. 26: Relationship between frequency of terms and the number of paragraphs in a set of data 89
Fig. 27: Hierarchical cluster of terms contained in the raw data set indexed by NetBase® 90

List of Tables

Table 1: Summary of Consumer Confidence Indexes from around the world……..…15
Table 2: A frequency count summarizing the Top terms & Top emotion words related to Capital Expenditure……...42
Table 3: The third output from a NetBase search; the indexed dataset extracted from social media networks for the search term of Capital Expenditure……..……44
Table 4: Illustration of the how niche social media networks could serve some subject area better than the others……...…78
Table 5: Indexed raw data from NetBase associated with Tax for 12 month period……..…85
Table 6: Part B of Table 5 showing the sampling interval…….……86

List of Acronyms

CBC	Conference Board of Canada
CC	Consumer Confidence
CCI	Consumer Confidence Index
CCS	Consumer Confidence Survey
IR	Information Retrieval
LIWC	Linguistic Inquiry & Word Count
MS Excel	Microsoft Excel
NLP	Natural Language Processing
PAYE	Pay-As-You-Earn
POS	Part-of-Speech
SM	Social Media
SPSS	Statistical Package for the Social Sciences
WSD	Word Sense Disambiguation

Chapter 1
Introduction

A notable indication of the economic prospects of an economy is the level of its consumers' confidence in the economy. In mainstream economic discussions, this is expressed as a quantified metric called the Consumer Confidence Index (CCI). It is a crucial data that it is referenced by businesses, governments and other institutions when they make announcements about the strategy direction their institution is taking or is about to pursue. In mainstream business media channels such as Bloomberg[1], time and again, the CCI is referenced and makes headline news.

The confidence consumers have in their ability to spend on goods and services is a cue of the direction of economic activities, whether they feel they have or will have the economic means to buy goods and services they require. These in turn influence the manufacturing of goods and delivery of services, affecting the levels of labor employment, the income tax generated for the government, profits or loss enjoyed by businesses and the commitment they go on to make on capital investment.

It could be argued that the historic reference to the CCI as an economic indicator has been time tested and thus is a 'reliable' way of telling whether economic activity will grow or decline in the near future. For example, the Conference Board of Canada has been publishing a CCI since 1980. During these past 35 years, despite the creation of other metrics to measure the consumers' confidence, the CCI is still referred to each time it is published.

Nonetheless, the reliability of the consumer confidence index, is directly linked to how accurately its methods of measurement capture the economic sentiment of the population of a country. In this regard, the methods of measurement start with a manual random survey of different demographics of the population. These surveys are known as Consumer Confidence Surveys (CCS) and they are issued by postal mail, electronic mail & telephone campaigns.

---

1 *Bloomberg News* of Nov. 28, 2014 illustrates the associated importance of consumer confidence in the article 'U.K. Consumer Confidence Stalls as Economic Outlook Deteriorates' (Ryan, J. - Bloomberg, 2014)

As retail technology has evolved, the measurement of consumer confidence has extended to including data from retailers' point-of-sales systems. This data is used to synthesize trends of consumers' buying pattern (Hudszinki, 2014), which is then used to infer the implicit confidence of consumers, based on how much they are spending relative to historic patterns. Complementarily, this information is combined with the conventional questionnaires, to give an indication of the general economic prospects. The recently published CCI of the United States was generated using probability design random sample (The Conference Board, 2014). It shows that consumer confidence stands at 88.7% for November, 2014 down by 5.1% compared to October,.

These variant but convergent methods of measurement tend to improve the accuracy of the CCI they are used to generate. On one hand some of these methods look to elicit the consumers' opinions on their economic situation—from the preceding six months to the six months ahead. This is the classic CCS method. Also, technology driven methods are used to monitor downstream spending patterns of consumers across vast retail horizontals. The diversification of the methods of CCI measurement can only have a positive influence because they entail looking at the subject of interest from more than one perspective.

When combined, these evolving means of measurement are destined to improve the accuracy of the CCI, from the researcher's point of view. Furthermore, following the advent of social media networks, new opportunities to generate more insights into the sentimental bias of consumers towards the economy continue to open up.

Using social media networks as a platform for social intelligence gathering is catching on as a data source in the market intelligence research industry. Companies such as Nielsen, Netbase, Frost & Sullivan, and Mintel enlist their use of content from social media networks, for intelligence gathering and marketing purposes (Netbase, 2010) (Nielsen, 2013). It is no surprise that a lot of research continues to concentrate on the improvement of analytical methods and frameworks to translate such content into useful information.

It is this translation that is pivotal to the usefulness and correctness of whatever information is synthesized from social media networks. For example, an area of study like *sentiment analysis frameworks* has originated concepts such as Natural Language Processing (NLP) that can be used as a tool to translate data from social media networks into meaningful insight.

In review, the indicator, CCI and how it is measured have expanded their scope beyond the use of questionnaires alone, and data sources present opportunities to gain another perspective on consumer behavior in terms of: economic habits (past), situation (present) or intentions (future). At this time, social media networks are a relevant sources for the above.

However, not a lot of work has been done to assess the characteristics of the data extracted from social media networks and the emerging trends reflected by and resulting from this data. Therefore, it is the goal of this research to explore at a macro level any and all trends that can be captured from the SMNs extracted data, using a commercially available social intelligence gathering tool.

The purpose of undertaking this research study is to explore the attributes of data from SMNs and assess how robust they are for the purpose of synthesizing a metric like CCI. This research would also help novice and established professionals in the area of electronic business technologies and data mining in particular to evaluate the issues and uncover the prospects of utilizing such data in research studies.

Therefore the scope of this thesis pertains to the use of a social intelligence gathering tool to extract data from social media networks. A qualitative analysis is performed on the data to evaluate such attributes as ‘representative’ of popular opinion during a period of interest. To do so, graphical word cloud summaries synthesized from the data are analyzed in order to see whether they give a hint about the economic sentiment of consumers.

In this introductory chapter, conventional methods used to collect data for the synthesis of a CCI are reviewed in Section 1.1. In Section 1.2, a brief illustration of how the CCI compares to other economic indicators is presented. In Section 1.3, social media networks and their attributes are briefly discussed, following the wave of research interest in them as a revealing source of the aspects of human psychological perspectives in the digital age.

Section 1.4 gives a detailed account of the objectives of this research and Section 1.5 lays out the methodology of research implementation. Finally, the structure of the remainder of the thesis is presented in sections 1.6 through 1.8.

1.1. An introduction to established methods of Consumer Confidence Index measurement

The conventional means of determining consumer confidence are different from one country to the other. Also, the organizations that provide this information are different, being either a not-for-profit think tank or government departments and sometimes profitable private businesses.

In Canada and the United States of America, telephone interviews and postal mail correspondences are used to survey a random selection of households. This could range from 2,000 to 5,000 correspondents (The Conference Board of Canada, 2014). The Conference Board which publishes the CCI for the United States posits five questions on the CCS questionnaire sent to corresponding households (McKinsey 2012). They are:

- Current business conditions?
- Business conditions for the next six months?
- Current employment conditions?
- Employment conditions for the next six months?
- Total family income for the next six months. Survey participants are asked to answer each question as "positive," "negative" or "neutral?"

The above questions are carried out at the following intervals: every two months and at the end of each quarter (The Conference Board, 2011). In recent times however, the Conference Board of Canada has partnered with Nielsen, a global market intelligence company in the retail space to provide its Consumer Confidence Index. According to the information on the CBC's website, Nielsen uses a technique known as probability design random sample on the primary data. Whether this primary data is solely based on conventional survey methods or data from other sources like social media networks is not explicit.

Other methods used to measure consumer confidence are predominantly those used for the survey methodology. Nonetheless, there is an increasing use of supplementary data from other sources such as retail point of sales and social media networks. In Section 1.2, the discussion is about how Consumer Confidence Index as an economic indicator is used.

1.2. Illustration of the use of Consumer Confidence Indices

CCI information influences business strategy decisions made by manufacturers, retailers, banks and governments, relative to other key economic indicators (Investopedia, 2014). A consistently declining trend in CCI would indicate that consumers have a negative economic

outlook which implies further decline in economic activity. While a positive CCI trend would be taken as antecedent to growth in economic activity in the near future—good for business.

For example, Fig. 1 attempts to show the correlation between different economic indicators in the United Kingdom and the level of the CCI over an unspecified period of time. This is one example of how businesses could base their decision of either investing in production capacity or cut labor costs, based on the anticipated economic activity inferred from the level of consumer confidence.

Similarly, banks can also take measures to curb the anticipated increase in borrowing by consumers and the use of credit cards. The government could come up with measures such as reduction of taxes in order to encourage spending. The opposite will apply in a case of an upward trend in the index (higher production/inventories and tax increase).

From 1921 to 1933, the United States government adopted a policy of laissez-faire which required them to be less involved in the affairs of the economy (however, there was no CCI at this time). They achieved this by keeping taxes low thereby encouraging businesses to invest more money for expansion and also creating more money for consumers to spend (BBC, 2014).

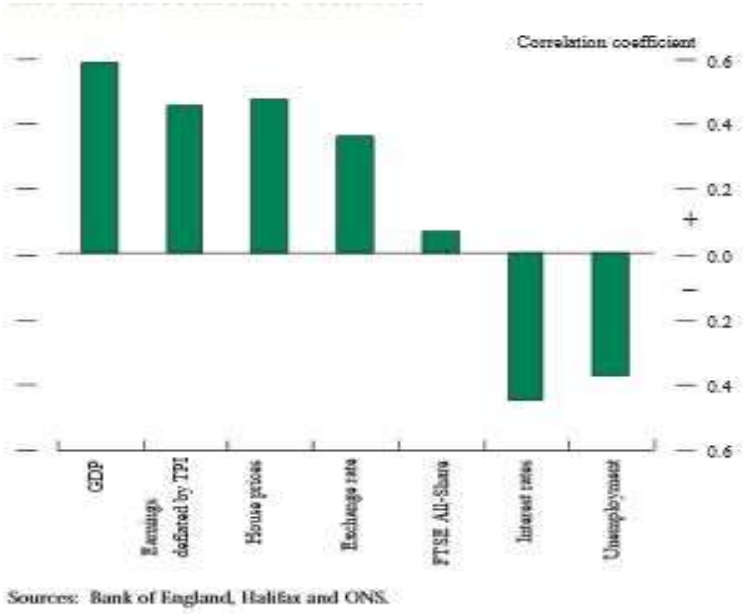


Fig. 1: A comparative illustration of the correlation between the CCI and other economic indicators

Therefore it seems reasonable to conclude that CCI influences decisions made by pivotal economic entities such as governments. One aspect of such dynamic has to do with the emergence of new sources of data that could be used to measure consumer confidence.

1.3. Social Media Networks as a source of data for Consumer Confidence measurement

In an era of social media networks facilitated by the advances of the internet, its components including blogs, Facebook®, Twitter®, FourSquare® et al, are a versatile source of information which could be harnessed for assessing the human perceptions that reflect their confidence in the economy. For example, Facebook and Twitter are popular networks used to express opinion, a state of mind or a perspective which could be mined for the sentiment of users relative to their economic situation. FourSquare is used to identify with geospatial preferences, such as where an individual is geographically located.

Also, blogs and the commentary they generate are a rich source of opinion bias towards a subject of discussion; e.g. an election, a new movie, a new government policy, stock market fluctuations, a commercial investment deal and much more. Aggregation of such information can be used to 'know' the economic state of consumers.

Furthermore, 60% of the world's population has access to the internet and now interacts using social media networks (Barry et al. 2011). Therefore tapping into this 'new' data source could give a supplementary boost to the conventional methods of data gathered to measure the level of consumer confidence in the economy.

How the richness of such data could benefit how information such as consumer confidence could be measured is one aspect this research aims to explore. In the next section the objective of the research is articulated.

1.4. Research Objective

Our research will explore first how representative past social media analytics data is/are as a source of information about the economic sentiment of consumers. Next, the study will analyze the intrinsic attributes of the data to assess their strengths and weaknesses relative to the level of consumer confidence in the economy.

1.5. Research Question

The research question is: *is data from social media networks suitable and reliable for measuring CCI?* Our research seeks to explore the attributes of data from social media networks to assess their robustness for use in the measurement of CCI.

1.6. Research Motivation

For the past ten years, electronic business technologies continued to transform the landscape of different industries, professions, politics, religion and the global economy. It is known that we are in a 'data-driven' economic era, where timely knowledge is vital.
This reality has prompted the emergence of new opportunities to apply e-business concepts in non-traditional areas. The measurement of CCI from social media networks is one such area.

The research student believes that by applying e-business tools and concepts, this research would facilitate an understanding of how to 'mine' data from unstructured sources like social media networks, develop further an ability to analyze such data and critically evaluate the usefulness of the data for application in a research area of interest.

The research will examine

i. How to use technological tools to exploit hidden insights in unstructured data from social media networks, in addition to the underlying concepts upon which such tools are built.
ii. The research would explore the versatility and suitability of social media networks data for measuring CCI.
iii. How to undertake data analysis and design models for macro information origination from data.
iv. Challenges of data gathering, data management and administration for research purposes.

1.7. Research Approach and Contribution

There are already tools that mine datasets for indicative trends such as *opinion or sentiment*. For example, General Inquirer from Harvard University, SAL—Sensitive Artificial Listener, TS—Twitter Sentiment, RA—RateItAll are examples of opinion mining retrieval tools (Tsytsarau & Palpanas, 2012). The limitation of these tools for purposes of this study are coming from the models upon which they are based, because the implicit assumptions may be unsuitable for the exploratory direction of this research. It is for this reason that a non-specialized SMNs data gathering tool was of interest. After studying available tools, NetBase® was identified as a suitable tool for this study, because its design is such that it can extract data from all SMNs that can be accessed through 'crawling' algorithms on the internet. It is not only

able to extract the exact conversations (Soundbites) from these SMNs but it deduces summaries about the data using various aspects of the NLP model it implements. It also makes the raw Soundbites data accessible for further analysis. The fact that NetBase® does not focus on the feed of a specific SMN and makes the extracted SMN data available for further analysis, is why it was selected for this research. It is recommended for other researchers who want to work with data from as many SMNs as possible, even though they will have no control over selection of specific SMNs.

NetBase® needs keywords as a guide to extract SMNs data that are relevant to the subject of interest. Therefore, the selection of the keywords was critical to minimize the extraction of Soundbites that are not connected to CCI. How it functions is illustrated in Fig. 2. To decide on the keywords to use, mainstream news media reactions and discussions were evaluated, after each publication of CCI in North America and Europe. Keywords such as 'capital investment', 'retirement', 'pension', and 'capital expenditure' identified on this basis.

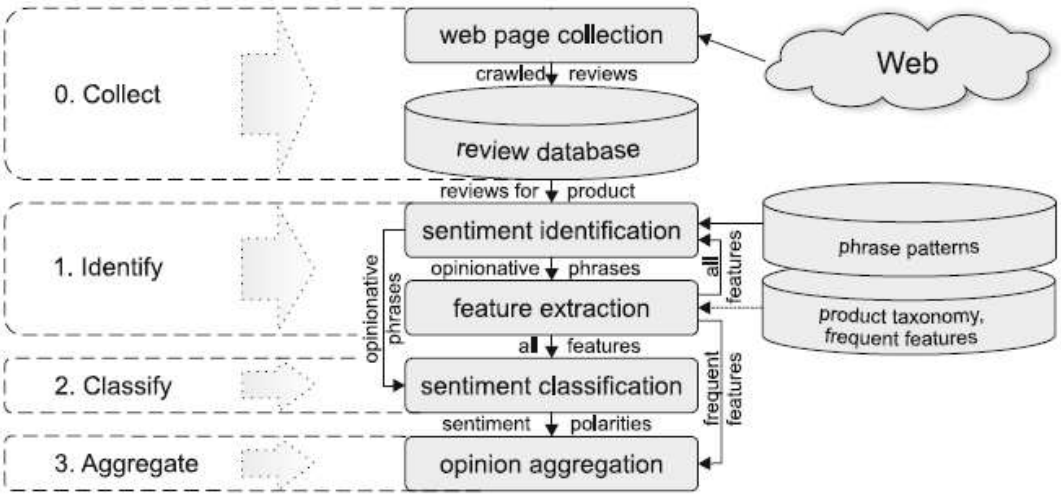


Fig. 2: Example of the architecture of a typical opinion mining application used for a product review (Tsytsarau and Palpanas 2012)

Noting however that these words were used in formal conversations, words would most likely be used in informal conversations, at a personal level, in response to CCI publications or events related to CCI were sought also. Words such as 'tax', 'health' and 'income' were identified from comments on SMNs like LinkedIn, Facebook, Twitter, Reddit and the webpages of mainstream news media houses; Bloomberg for example.

Identifying these words were one part of the challenge, understanding which words were likely to be used in most conversations on SMNs was the other part of the challenge, no easy method exists for. So we had to turn to other works in the literature, which had to devise some

nomenclature to mine data from SMNs. Then the syllable of these identified words were noted: in private conversations single syllable words are aptly adopted than words with more than two syllables. Since there were no one tools that could be used to justify this observation using data from SMNs, another perspective was sought from other related work in the literature (Daas and Puts 2014). A combination of these perspectives and consultation with supervisors led to the selection of fourteen keywords used to in NetBase to extract data.

Research Contribution: There are claims in the literature that data from SMNs are noisy. In order words their content are found be contain more irrelevant data than useful data, albeit for the particular research purposes. Since there are no obvious attempts on how to quantify this noise, one of the techniques used in this thesis to assess the *representativeness* of SMNs data can be reused by researchers to gauge how well their mining tool, or keywords are performing relative to a subject of interest. The method is simple: it uses the SMN with the highest number of users, Facebook, as a benchmark. Then assuming the data mining tool aggregates summaries like NetBase does, the volume of the most used word in the 'top term' is divided by the benchmark, and multiplied by 100. If this number is very small < 1%, it would imply that a small proportion of SMNs data are directly related to the subject of interest. This metric can easily be calculated for each top word under the results obtained per keyword. Such that, if the metric is small, it indicates that the keyword may not be adequate and so could be substituted with another keyword. And the process can be repeated until this ratio becomes >1%, for each of the keywords used to mine data. Thereby improving the representativeness of data got from SMNs and in turn the accuracy of the information sought.

1.8. The structure of thesis

It is necessary to understand the ongoing effort to extend the body of knowledge to the context of this research study. Concepts have been postulated about the relevance of the CCI and the significance of variables used in its computation. These concepts and research applications published in the literature are reviewed in Chapter 2. Chapter 3 describes the methodology of the research in-depth. It will cover aspects that have to do with the means through which the end goal will be achieved. Chapter 4 deals with the processing of data to yield exploratory insights into the richness of such data for the purposes of understanding patterns that relate to the confidence of consumers in the economy and to know whether or not social media is a good tool in measuring sentiments. In Chapter 5, the results of Chapter 4 are discussed to highlight the learning outcomes; then a conclusion and recommendation for future research studies are given.

Chapter 2

Literature Review

The use of data from social media networks to synthesize macro information continues to elicit a tremendous amount of research interest. There is a huge wave of research applications employing data from social media networks to compute informative metrics, such as the level of customers' satisfaction relative to quality of business service delivery. An example of this is ratings that show the punctuality of train services based on customer's complaints or commendations on social media networks.

The first part of the literature review focuses on the concept of consumer confidence as an economic indicator.

The second part concentrates on the tools and concept of data mining. This is important since these are the interfaces through which data from social media networks could be extracted.

The third part of the literature review will explore whether data from social media networks contains attributes that could give an exploratory perspective on variables that have been identified as components of the consumer confidence index indicator.

2.1. The economic indicator of consumer confidence index

The publication of a Consumer Confidence Index started in 1946. It was based on the University of Michigan's Survey of Consumers. Since then, different aspects of its versions have evolved. For example, today's version includes derivatives like the frequency of its measurement, its sample size and method of data collection, including the dependent variables and independent variables, the rules governing whether respondents from a previous data collection exercise could participate again, and what the maximum percentage of such respondents would be are added (Merkle, Daniel M; Langer, Gary E; Sussman, Dalia;, 2004).

It has also been suggested that the CCI was established to produce an economic indicator that would 'lead' imminent economic activity and thus be used to forecast consumer expenditure (Roberts & Simon, 2001).

When it comes to how the CCI is determined, three of the oldest forms of measurement constitute the foundation upon which today's concepts are built. They are ABC News/Money,

the Conference Board (of the USA) and the University of Michigan consumer confidence surveys. Pertinent information about their methods is shown in Fig.3 (Roberts & Simon, 2001).

All methods shown in Figure 3 are based on the assumption that *sentiment* can be used to predict household spending, especially for durable products which are 'discretionary' in nature, given the ease with which their purchase could be postponed. As Mueller (1963) stated, "confidence variables were close predictors of durable and non-durable household expenditures." (pp. 899-917).

Methodology of Consumer Confidence Surveys

	ABC News/Money	Conference Board	Univ. of Michigan
Method	Telephone	Mail	Conference
Sampling	RDD with random selection in household	Selection from a non-random panel	RDD with random selection in household
Weighting	For probability of selection and to Census (region, age, race, sex and education	Not disclosed	For probability of selection and to Census (age and income)
Sample size	About 1,000 (250 per week x 4 weeks)	About 2,500 for end-of-month release; 3,500 for later revision	250-300 for mid-month release; 500 for end-of-month revision
Field period	Wed-Sun each week; Results based on a four-week rolling average	Sent first of the month; Accepts returns through end of month	Around first of the month through a few days before the release
Release	Weekly, Tuesday evening	Prelim. Figures, last Tuesday of month; final figures with next month's release	Preliminary figures at mid-month; final figures at end of the month
History	Started in December 1985	Started bimonthly in 1967; went to monthly in 1977	Started annually in 1946; quarterly in 1952 and monthly in 1978

Fig. 3: Differences between the three oldest forms of consumer confidence index measurement (Source: Reserve Bank of Australia)

It has been noted that consumer survey results provide confidence indicators and important information on assessment by each consumer concerning their economic situation and their various expectations for the near future difficulties in application continue however. For example according to Mueller, the Michigan Consumer confidence survey, an important predictors of consumer expenditure on durable and non-durable goods was unable to ascertain the future performance of the economy due to its lack of confidence variables at the time (Mueller, 1963).

In 2006, Cotsomitis and Kwan demonstrated how they were able to forecast the expenditure of households by combining two economic indicators. The first was the CCI which indicates the economic prospects of the economy by measuring the level of consumers' spending. The second indicator is the Economic Sentiment indicator which identifies the incentive(s) behind the consumers' spending activities. This slightly differs from the approach of Katona (1960), who claimed that the best way to measure the [economic] confidence of consumers is through their discretionary expenditure.

Research has extended into components of the different methodological strategies used by these three consumer confidence indexes. Chief among these components are the sample size, the weightings used, the sampling and the method of data acquisition, such as telephone or mail.

All of these components continue to drive the evolution of the consumer confidence index which originates from consumer confidence surveys.

There is, however one fundamental limitation of survey based methods. It has to do with the non-participation of respondents and this is deemed as a higher risk due to inaccuracy compared to the risk of unrepresentative information that could result from the use of a small sample size (Roberts & Simon, 2001). This is the case of the Conference Board of Canada. On its website it is stated that the size of its sampled respondents is 2500 – 5000, but only about 1,500 respondents replied to the survey (The Conference Board, 2011) (Bloomberg News, April 2003).

This limitation continues to spur research efforts to clarify and develop evidence on associations that form the core of the economic indicator of consumer confidence. One such study examines the association between the variables of sentiment and economic activity, identified and discussed in the American literature. Nonetheless there is little evidence on the strength of this association, why it exists and its causal direction (Roberts & Simon, 2001).

Yet in another regard, the index of consumer confidence is subjected to rigorous analysis, especially when associations with economic activity are being discussed. Whereas some national indexes of consumer confidence are derived by weighted or unweighted contributions of other indices, which do not get a fraction of such critical analysis.

For example the University of Michigan's CCI is determined using the contribution of the index of Consumer Survey, a little known survey (Merkle, Daniel M; Langer, Gary E; Sussman, Dalia;, 2004).

The majority of previous research focused on the economic indicators of consumer confidence, and refers to evaluating how the different components used to determine it are used, ignored, assumed to be significant or modeled. In particular, the choice of variables and their assignment as being dependent or independent for the application of statistical procedures like regression is considered. This continuous variation leads to a continuous revision of the consumer confidence index even after it has been published[2].

In summary, generating an index for consumer confidence is based on a method of data collection that has not changed much since 1946. Instead, it is the methods of processing the data, how the sample of the population that participates is rotated, the rules governing the participation of previous participants in new surveys and the data models used to develop an interpretation from the data that has gained widespread attention in the research community.

Considering the above, it is evident that the composite aspects of the consumer confidence index, i.e. other indices like Consumer Sentiment used to determine the University of Michigan's consumer confidence indices, play a pivotal role in the indices representativeness. Also, these other indices relate to aspects of human sentiment or emotion that could be traced to and extracted from human interactions.

For example, in May 2014 the *survey of consumers* by Thomson Reuters and the University of Michigan found that respondents expressed concerns about slow wage growth. This would be indicative of a decline in living standards for households in the year ahead, after their income is adjusted for inflation. In another context, respondents hinted at hesitant buying intentions, citing fewer discounts on prices and higher interest rates (Thomson Reuters & University of Michigan, 2014). This leads to the hypothesis that if people are happy to share the realities of their economic situations in surveys, they are likely to do the same on social media networks.

In Fig. 4, the National Restaurant Association was able to survey its customers to understand their personal finance and they were willing to share this information. The outcome was described by the following statement, "Nearly six in 10 adults say their household finances are either in fair or poor condition, which is essentially unchanged from their appraisal of

[2] (The Wall Street Journal, 2014) (FxPro Blog, 2014) (Reuters, 2014)

their finances in 2010, when the economy was in the early stages of recovery" (National Restaurant Association, 2014).

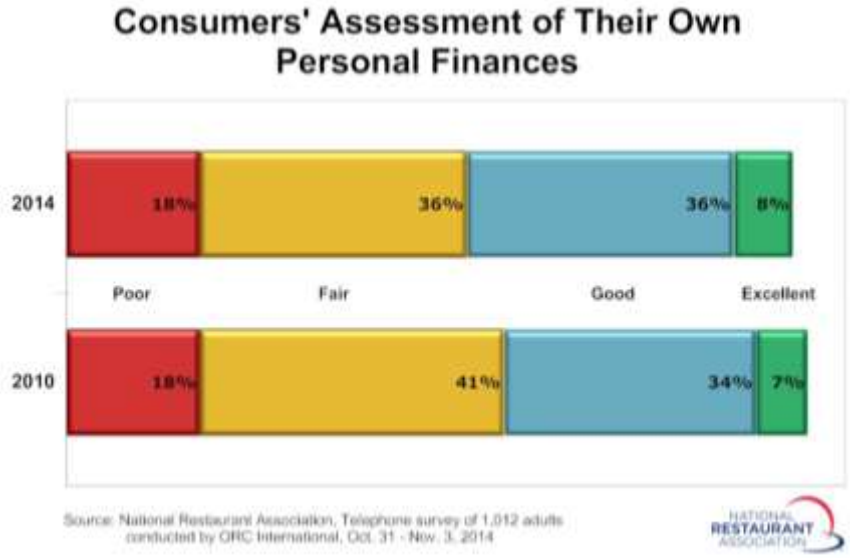


Fig. 4: A survey of restaurant customers on their assessment of their personal finances

Clearly, this presents an opportunity to use sources of information where human interaction is the exchange currency—and this source could potentially be social media networks. Since the broad goal of this research is exploratory, it is worth attempting to show that data from social media networks do have and reflect at least a few representative similarities with the information obtainable from composite indices of the CCI.

But to do so will require a look at what transpired from the body of published knowledge that continues to shape how associative models are used to link human interactions with various subjects of interest. The body of knowledge that has propelled this interface of information extraction will be reviewed next in subsection 2.2.

For completeness, a review summary is given in Table 1, where consumer confidence indices from around the world are explained in terms of their methodology and approaches.

Table 1: Summary of Consumer Confidence Indexes from around the world

	Canada	United States of America	United Kingdom	France	Nigeria
Organization Responsible	Conference Board	Conference Board and university of Michigan consumer sentiment Index	GFK and MORI.	Commission in the Journal European Economy (Supplement B).	MasterCard worldwide index.
History	The conference board started its operation in 1980 and has been an independent research organization since inception. (Conferenceboardcanada 2014).	The Conference board started in 1967 and Michigan's Consumer index started in 1946.	GFK started across Europe as early as 1970 but officially started in UK in the year 1995.		
Method	Telephone	The conference board uses mail to acquire its own consumer confidence data while the Michigan uses Telephone as its method of acquiring the data.		Telephone.	Gathering information based on usage of master cards by the consumer.
Size	Approximately 2,000 households are used for the survey.			The survey is carried out on approximately 2000 households.	
Timing and Data Release	In 1980 the consumer confidence data was produced once in four months, in 2002 the pattern was changed and the data was then produced monthly.	Both organizations starts carrying out their survey first day of each month. The conference board releases its result at the end of the month whereas Michigan publishes its initial result on the second Friday of the month which is around 50 – 60% of the total number of response collected. The final results are published at the end of the month (Curtin 2003).	The survey is carried out and published once every month. GFK focuses its questions on household finances and the general economic outlook.	This telephone numbers are selected randomly and dialed every first three weeks of the month except in August. July and September are used to calculate August.	This survey is carried out twice every year in the various geographical region in the country. (Vanguardngr 2011).

Operations and Questions	The consumer confidence survey is bases on four specific questions: • Considering everything, would you say that your family is better or worse off financially than 6 months ago? • Again, consider everything, do you think that your family will be better off, the same or worse off financially 6 months from now? • How do you feel the job situation and overall employment will be in this community 6 months from now? • Do you think that right now is a good or bad time for the average person to make major outlay for items such as a home, car or other major item?	• The indices produced by both Michigan and Consumer board based on questions on the current condition of the economy and questions about future expectations. • For the measurement of the current condition the consumer board questions are only directed towards the view of economic conditions and not the personal experience of the consumer whereas Michigan questions are directed towards family finances and the peoples spending in general. • In order to effectively measure future expectations both the conference board and Michigan assign two question directed specifically towards the economic condition and one question directed towards the consumer's personal finances. • One of the major differences is the structure of the question, Consumer board tends to ask questions about the consumers view of the condition in their local area but Michigan aims at gathering data based on the customer views about the condition of the nation as a whole rather than their local environment. The approach of questions used by these bodies was widely criticized by Dominitz and Manski (2004).	GFK survey in the United Kingdom is based on five specific questions. They include: • How has the financial situation of your household changed over the last twelve months? • How do you expect the financial position of your household to change over the next twelve months? • How do you think the general economic situation in this country has changed over the last twelve months? • How do you expect the general economic situation in this country to develop over the next twelve months? • In view of the general economic situation, do you think now is the right time for people to make major purchases such as furniture or electrical goods?	The questions asked are based on five indicators, they include: • Personal financial position –Past change • Personal financial position- outlook • Likelihood of major purchases • Living standards in France- past change • Living standards in France- outlook	MasterCard index worldwide survey questions in Nigeria are based on the following sectors • The Economy as a whole • The rate of employment in the country • The stock market • Regular income of the consumer • Quality of life experience or expected by the consumer

Computation of Indices (Positive and Negative)	After collection of consumer confidence data, the positive and negatives are used to compute and analyze the data in order to provide a result to the general public on the consumer confidence of the country's economy. When the positive and negative has been sorted out the index is then computed for each question using the formula: *Percentage of positive responses/ (percentage of positive responses + percentage of negative response)*	According to Daniel M., Gary E., Dalia S., *"The Conference Board index is computed by taking the positive percentage for each question divided by the sum of the positive and negative percentages. This number is then divided by the base year value from 1985 and multiplied by 100. The resulting values from each question are averaged to form the overall index"* *"Michigan computes its index by taking the difference between the positive and negative percentages for each question and then adding 100 to each. These are summed and then divided by a factor representing the base year, 1966. Finally a small correction is made to account for a design change in the 1950s".*	The indices are computed using the positive and negative values. According to Stuart B., Melissa D; *"The results are published as a net balance of positive less negative responses, with those who answered 'a lot' in either direction given twice the weight of those who answered 'a little'".*	The consumer confidence is the calculated using positive and negative responses with formula; *% of positive response - % of negative responses.*	The index is computed using positive, negative and neutral value. 0 values is assigned as the negative value which is known as the most pessimistic, 100 is seen as the most positive value which as represented as the most pessimistic and 50 is seen as the neutral value.

2.3. Measurement of Consumer Confidence

The management consultancy firm McKinsey, published a report entitled "What executives think about the economy (2014)" – that aimed at deducting the confidence of business executives relative to the confidence reflected by consumers in their respective geographic business locations globally (McKinsey&Company,2012).

Nielsen Global Group is another organization that measures consumer confidence since 2005. From across Africa, Asia, Europe, Middle East, North America and Latin America 30,000 respondents are surveyed in order to quantify their economic sentiment/outlook in the 'economic' segments of their lives.

The Nielsen Group report revealed that the global CCS data collated in the last quarter of 2013 shows that CCI has remained the same with an index of 94 points (Fig. 6) for three preceding quarters unlike the year before which was at a 91 index point. This means that the global consumer confidence has increased since 2012.

Although this CCI figure is for all the countries involved, an increase in the global index data is not representative of an increase in respective individual countries. The last survey carried out on the 29th November 2013 was the fourth quarter in 2013, and it showed a CC increase of 43% of the markets which was lower compared to the 57% increase which was measured in the third quarter of 2013. The highest consumer increase was seen in Indonesia with a point of 124 whereas the lowest decrease was seen in Portugal, Italy, Croatia and Slovenia with a declining index point of 44 (Nielsen, 2013).

We can see that the low CC of Portugal and Italy, for example, correlates to the findings of the World Economic Situation and Prospects report (United Nations 2015), that indicated the weakening of these economies owing to a slowdown of export driven business in 2011 and 2012 (United Nations, 2015). In turn this slow down impacted the profitability of local businesses that enjoy significant export business, thus leading to less capital investment, cost cutting measures and other cash flow sustenance measures that further impact the economic situation of the workforce who is also the primary consumer source. Thus it is these impacts that Nielsen has managed to elicit through its surveys.

In Fig. 5, a hand count of the number of countries that experienced an increase in CC shows there are 26 compared to the 32 with a decreased level of CC and 2 countries that experienced no change. As such it is evident that the CCI measured in the preceding three quarters, i.e. January to October 2013 must be a net increased CC around the world.

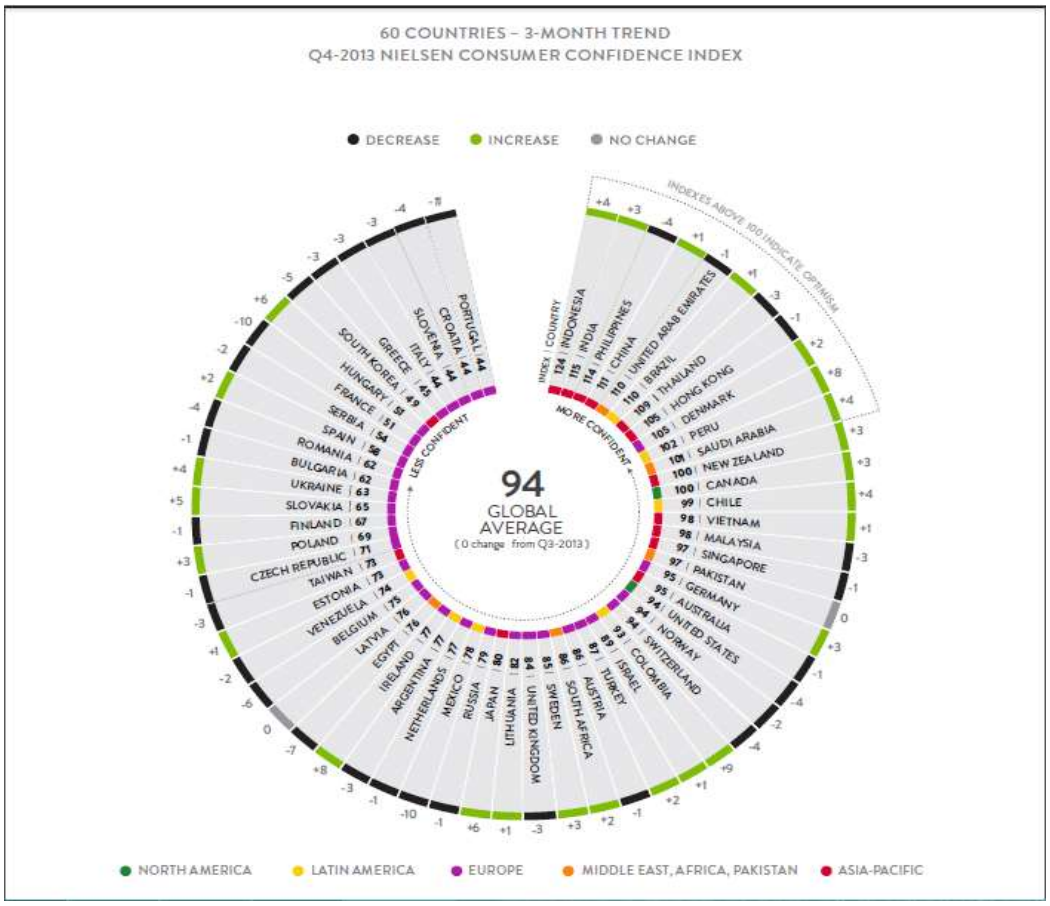


Fig. 5: A global CCI of sorts generated from a survey of 60 countries for the period of October – December 2013; Source: Nielsen, Consumer Confidence Concerns and Spending Intentions around the World Quarter 4, 2013.

2.4. Social media networks

Social Media is composed of online technology tools which enable communication between different people from around the world via the internet.

According to Margaret Rouse 2011,

> *"Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration."*

For use in the economic and business sector, Social media involves the gathering of comments made in online forums, created with the sole purpose of educating people about products, brands, services and general issues in the economy (Blackshaw & Nazzaro, 2004).

Magmold and Faulds (2009) defined social media as a wide range of word-of-mouth forums which allows users to be able to express their feeling about a topic or even create a topic for discussion. The forums included: social networking sites, blogs for personal and company discussions and product rating sites.

There are different types of social media in the world today, they include for instance, Facebook, Twitter and LinkedIn. The communicating parties are able to connect and hold conversations using formats such as; text, audio, video, images and other multimedia.

Social media has become one of the strongest and most powerful tools used online since it allows thousands of users to be reached within the shortest minimum time possible. According to statistics provided by socialnomics.net, on the speed of development of the social media, it took the radio 38 years to reach 50 million users, TV 13 years to reach 50 million users, but it took Facebook just 9 months to reach 100 million users.

The tremendous difference in outreach suggests that social media is an extremely fast way to get to customers or users about a particular product or issue. Social media has not only been used for personal use but it has been used for business as well. Using social media has become one of the new methods used by The Economist to gather data on the level of consumer confidence in the economy, the methods including: counting the number of likes and dislikes of a topic, how many comments have been given by consumers on relating topics and gathering how many re-tweets were done by consumers.

This method of gathering data shows the real consumer sentiments towards a brand or the economy based on the comment provided by the consumer. According to a report from Performics, 2010, a survey carried out on 3,000 U.S social media users revealed that social media has a tremendous effect on customers attitude towards brands. The report also showed that 53% of the authenticated Twitter users recommended a company and its products by tweeting about it and 48% of the people actually delivered on their intentions to buy that particular product (Performics, 2010). According

to Keller Fay Group in 2010, the average consumer in a country mentions his/her specific favorite brand approximately 60 times in a conversation with family, friends and on social media.

SMNs are rich with expressions of sentiment which is important to the measurement of CCI. *Sentiment* is known as the feeling or reason for an expression made behind a comment and reference about a particular topic or brand. Various tools have been made which enable measuring the consumers' sentiment. From a business perspective, the ability of a business owner or a brand to be able to measure or understand the consumer's sentiment behind a comment is very beneficial to the growth of the organization. This helps to determine the state of mind a consumer was in as at the time he/she made the comment.

Research was carried out on how to effectively predict movie sales from bloggers sentiments (Mishne & Glance, 2006). The researchers stated that the volume of discussion about a product on the social media especially web blogs, cannot guarantee a product's final performance, therefore they advised that sentiments on the weblog data should be analyzed as well as the volume of the data, concluding that positive sentiment - which is different from positive result- is actually a better data for the prediction of the success of the movie, rather than just the positive comments.

A similar analysis was carried out on how social media can help predict the future outcome of a new movie before it is released. In this research, Twitter.com was used to forecast box-office revenues for movies instead of the static blogs which were used by Mishne and Glance. The result from the research was calculated by analyzing the number of tweets that are currently on the network based on a particular movie.

These tweets were divided into positive and negative tweets, Positive tweets are interpreted as representative of customers looking forward to watching the movie based on the preview, while the negative tweets mean that the customers are not interested in watching the movie. A linear regression model for predicting box-office revenues of movies in advance of their release was constructed. The results showed that there is a strong correlation between the amount of attention or discussions consumers have about a movie about to be released and its subsequent ranking when the movie is eventually released (Asur & Huberman, 2010).

Looking at the two research studies it can be concluded that the result derived from the dynamic method (tweets) can be used by itself to forecast the revenue to be made from the forthcoming

movie whereas the result derived from the static method is not enough to predict movie sales, the sentiments behind the content of the blog must also be considered.

Chris Barry, Rob Markey, Eric Almquist and Chris Brahm (2011) used a Net Promoter system to measure how the social media has affected the consumers' confidence on a product or the organization as a whole. The questions asked on the social media about the product included "How likely would you recommend [this company or product] to a friend or colleague in social media?" After analysis was carried out, they stated that:

> *"Customers who engage with companies over social media are more loyal and they spend up to 40 percent more with those companies than other customers."*

In 2012, Schweidel, Moe, and Boudreaux carried out an analysis of the potential to get *brand sentiment* from social media conversations. They used data which was collected from different social media domains such as Facebook and Twitter. They proposed the use of a hierarchical Bayesian regression model which was used to measure these sentiments effectively.

Based on the research carried out, it has been discovered that social media is one of the fastest and easiest mediums that can be used to get large amounts of consumers reviews which can be used eventually to calculate the overall consumer confidence on a brand or the economy as a whole. Different approaches on what actually affects consumer confidence were also researched. Among them are unemployment, rising inflation and stock price, but the greatest question still remains "What actually causes a rise and fall in consumer confidence?"

In conclusion, the social media itself can actually cause a rise and fall in consumer confidence in the world today as information about the economy can easily be known by consumers as quickly as it happens (using blogs, Facebook, twitter). This affects how consumers feel about the economy unlike the days when such platform was not easily accessible.

2.5. Attributes of Social Media Networks

Social media in recent years is large and well known for its social network and content sharing abilities—videos and pictures (Asur & Huberman, 2010). It can also be seen as a setting for virtual discourse where people create content, share it, bookmark it and network at an extraordinary

degree and speed. It is also used as a real-time snapshot of updates of interest, location, instant memories and so on.

The use of social media networks has brought about one of the biggest cultural shifts since the industrial revolution, attracting up to 1.82 billion users globally (Statista, 2014). It is engendering a whole new wave of human interaction, different from the structured methods that existed before it. The increased use of mobile internet technology has also contributed to the growth in adoption of SMNs illustrated in Fig. 6. Some examples of popular SMNs as seen in the figure include: Facebook, Twitter, MySpace, Digg, Reddit, Instagram, foursquare and sound cloud.

Social media comes in various forms:

- **Wikis**: This act is done by adding articles and editing existing articles. Example; Wikipedia, Wikia.
- **Social News**: This act is done by voting for articles and commenting on them. Example; Digg, Propeller, Reddit.
- **Social Networking:** This act is done by adding friends, commenting on profiles, joining groups and having discussions. An example would be Facebook, Hi5, Last.FM
- **Social Photo and Video Sharing**: This act is done by sharing photos or videos and commenting on user submissions. Example; YouTube, Flickr.
- **Recruitment:** This act is done adding suggestive jobs available, commenting on groups formed by companies, sharing various job experience and also applying for jobs. It is the main source for e-commerce recruitment. Example; LinkedIn.
- **Social Bookmarking**: Example is Del.icio.us, Blinklist, Simpy. This is used by tagging websites and searching through websites bookmarked by other people.

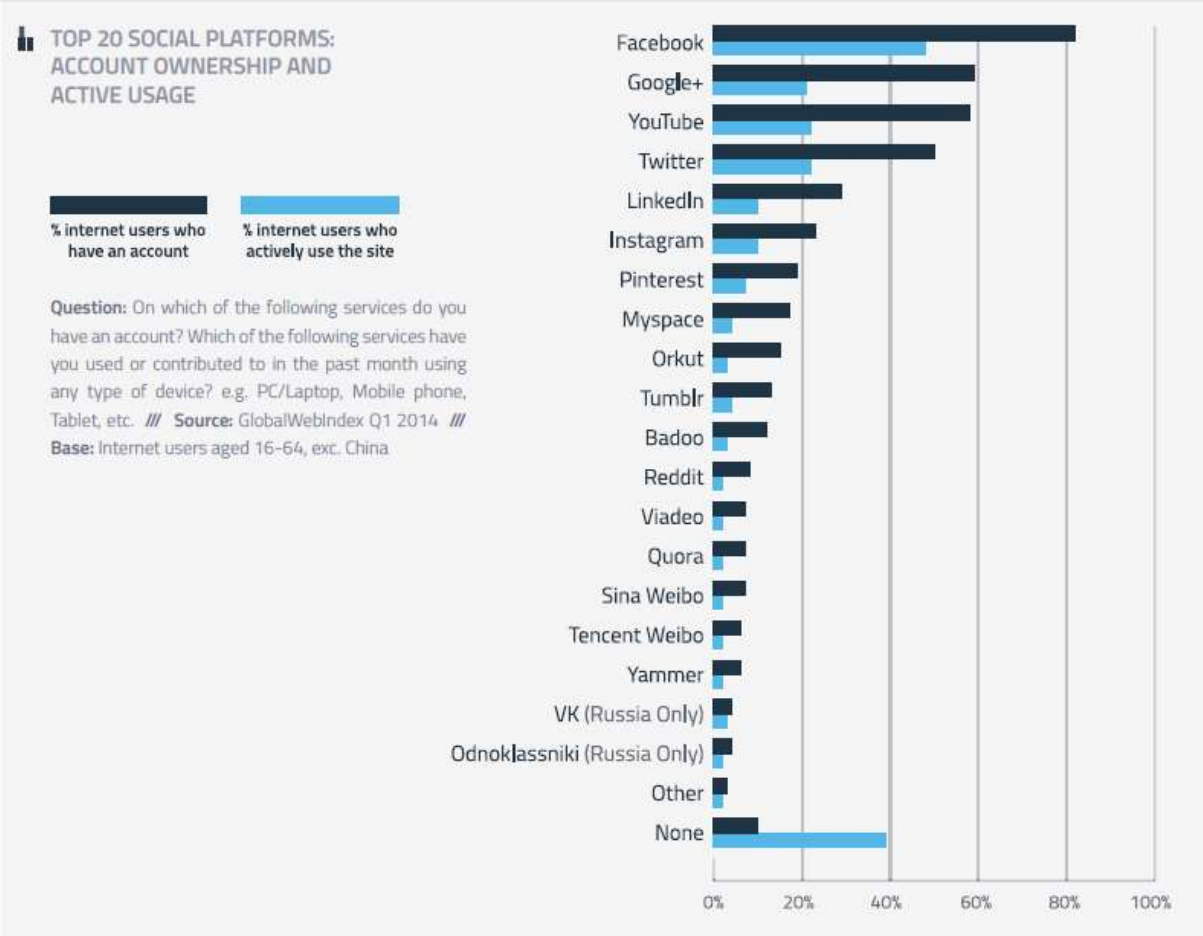


Fig. 6: Global Penetration of social platform-Account ownership and Active usage— (Chaffey, Dave, 2014)

2.6. Data mining concepts

The mainstream approach of generating the economic indicator of consumer confidence is through surveys and it has proven its reliability over time (Merkle, et al., 2004). Nonetheless the process of generating this index is becoming hybrid, where traditional survey methods of mail or telephone calls are used to gather data and the information extraction models are used to extract additional data from other [electronic] sources. This is a new approach that is gathering momentum, especially the use of social media networks. The Conference Board of the U.S recently outsourced the creation of its consumer confidence index product to Nielsen who uses this hybrid approach (The Conference Board, 2011).

Also, there is an uptake in the use of data from social media networks for the generation of insights into multiple areas of interest related to human interaction. At the core of this process are models that aim to associate the textual content of social media networks with the intentions of the

interacting parties. This has driven research efforts in the area of information retrieval, commonly known as IR.

Information Retrieval (IR) methods have been developed to optimize the extraction of insights from data in social media networks. NLP and LIWC are the well-known examples. In the financial services industry for instance, innovative companies such as Barclays, TD Canada Trust and the Bank of America leverage such data to succeed in very difficult sectors like pay day loans, a banking business segment whose clients are known to default often.

Wonga® a payday loan company in the UK, pioneered the use of data from social media networks to mitigate the industry-wide challenge of defaults. It uses up to 8000 data points automatically captured from social media networks to augment the profile of borrowers based on information they have provided. This data is then aggregated to classify whether the prospective borrower will default or not. It is claimed that this application of information from social media networks reduced the percentage of defaults by 50%—a record breaking feat going by the industry's standards (Tobin, 2012).

In light of the fact that the storage of human communications is in text format, models of lingual origin continue to be developed to correctly associate *meaning, context and intention* expressed in such pieces of information. In fact this remains the challenge plaguing various frameworks that have been proposed in the body of literature.

It turns out that in order to retrieve information from a textual source, word content must be linguistically classified, after which they are associated to a segment of human life (e.g. emotion, religion, social, economy and cultural). Then they are aggregated to piece together the *intent, context or meaning* of the interaction held by the 'writer(s)' and the 'listener(s)', as it may be. The reality is that there is no one framework that captures all three correctly.

Frameworks have therefore been targeted at either deriving *meaning or the intention or context* of a communication made known on social media networks in particular. A good example is a framework that uses *argumentative* filtering to establish whether a piece of information expresses a negative or positive point of view/intent. Other frameworks apply concepts such as NLP or LIWC. This research study simply aims to investigate if the attributes of data extracted using such

micro level models is good enough to synthesize macro information such as the generation of a CCI.

Although there is a high proportion of visual-audio content in some social media networks such as YouTube™, these are excluded from the scope of this research. Based on observation, social media interactions predominantly involve text based communication. The majority of users would much rather write a response or comment, instead of posting a video or audio response. This is the case when matters such as government policy, labor rates, unemployment, tax increases or the economy in general are being discussed, in contrast with the visual media which is more of a social interaction tool for marketing and other forms of 'soft' topical interactions. Of course, most executives, organizations and governments post videos to engineer engagement when discussing matters of a serious nature. But this is a small number compared to the majority of social media network users, who also play the role of consumers in the economy.

There are a variety of concepts that postulate models for mining textual content for information. In some instances they have been used to predict the sales of movies or even predict the outcome of elections. The U.S presidential election won by President Barack Obama in 2008 is a well-known example of predicting wins (INSEAD, 2009).

2.6.1. Natural Language Processing concept

NLP is a branch of computer science that deals with the processing of human language so that it is accessible to computer applications, in such a way that it enables applications to process, classify, aggregate and generate insightful information of a higher human understanding (Microsoft Research, 2014). The earliest efforts of NLP application started in the 1950s and since then NLP has been behind a diverse number of applications. Examples of such applications include: information extraction, machine translation, sentiment analysis and question answering (Research at Google, Google Inc., 2014).

These have been materialized into products like the voice recognition applications for smart phones and cars used to identify the task their user intends to perform, spam detection in electronic email, the Online web chat used by companies for web based customer support services, the IBM Watson computer and so many more (Jurafsky & Manning, 2014) (Research at Google, Google Inc., 2014) (IBM, 2014).

NLP is a fusion of a long list of tasks that solve a piece of the language processing puzzle, as the incentive to reduce the latency of language translation for computer processing grows strong. Some of these tasks have led to solutions of a basic level of language processing. For example, Name Entity Resolution which has to do with identifying words in a text that refers to people, places or institutions, and Part-of-Speech (POS) tagging where a word in a text is identified as being either a noun, verb or pronoun (Collobert, et al., 2011).

NLP uses semantic rules at different levels to associate components of language. For example, it analyzes nouns and phrases, tagging each to label it as either a person or organization and clustering a collection of these components so that those that refer to the same entity are easily classifiable.

The depth of this breakdown of language based on rules depends on whether the source of the data is closed or open and unstructured like the internet, Web 2.0 (Research at Google, Google Inc., 2014). Historically, most NLP research has focused on the former and only few companies like Google are advancing the state-of-the-art of NLP, where the text being processed is varied and does not conform to the structure of spoken language.

Similar to the fundamental constructs of language, NLP employs the syntactic rules of structured language to predict part-of-speech tags for each word contained in a sentence or document being analyzed. This sort of 'parsing' has proven to be a fundamental challenge for modern day applications of NLP, which have to deal with multiple human languages without any communal lingual structure.

Other tasks are aimed at tackling language processing challenges that are at an intermediate level and steady progress is being made. Examples of these include Sentiment Analysis where the goal is to extract subjective information from a sentence or body of text and thus determine its "polarity". Polarity is an indication of whether it is for example positive, negative, in-support and against, depending on the use case objective. Others include Information Extraction, e.g. the extraction of a calendar date from an email invitation specifying the date and time of an event. Then there is Word Sense Disambiguation (WSD) and Machine Translation to name a few (Jurafsky & Manning, 2014).

Finally there are tasks that have to do with macro information synthesis. These border on the production of insights that are deeper than the lateral interpretation of language. For example, in the summarization section we have the following sentences (i) the London Exchange is up (ii) the Standards & Poor 1000 index rose and (iii) employment jumped, the summary that could be deduced is that the economy is good. This is a herculean task for a computer to undertake, because it needs more than just semantic and syntactic rules of language to reach such insight. It needs knowledge and experience equivalent to that of an individual that is aware of the existence of S&P, the London Exchange and the fact that they are strongly associated to the state of an economy. Other tasks in this category are Paraphrasing, Question Answering (QA) and Dialog for web help services (Kongthon, et al., 2009) (Jurafsky & Manning, 2014).

NLP is a concept that has enjoyed considerable research and advancement. An aspect of NLP that could be applied for this research is Sentiment Analysis and there are other aspects such as Parsing, Summarization, POS and WSD.

Assuming that the majority of the data on social media networks follow language rules, NLP is one approach that would enable the extraction of relevant classes of conversations, particular words and possibly the polarity of conversations relative to the economic situation of the conversationalists. Therefore, the selection of a tool that would be used for this research would be anchored on whether it implements NLP algorithmically.

2.6.2. The Linguistic Inquiry and Word Count concept

The Linguistic Inquiry and Word Count (LIWC) methodology, developed by Pennebaker, Booth and Francis in 2007, was designed specifically for analyzing text in order to measure the emotional expression it contains. Since this research is being undertaken at a macro level, the LIWC concept is not the right tool. Another approach is the Life History Calendar. This concept is based on cyclical occurrences in life, such as birthdays, weddings, childbirths among others. It is believed that an expression of one's disposition towards these events would give an indication of their situation at that point in time.

2.6.3. Life History Calendar (LHC)

It is a concept that is used retrospectively on data about events occurring in people's lives. Such as: when they got married, graduation from or enrolment into a new education level and first job.

In 1969 a national sample study was carried out in the United States using the LHC technique. In this study, various kinds of data such as: timing of education, employment and family events were recorded through the use of the technique (Freedman et al, 2011). The LHC concept is biased towards seasonality, an attribute which constitutes a small part of conversations on social media networks. For this reason it does not represent a useful approach for this study.

LHC has two basic components:

- A visual feedback of answers which includes multiple life domains; examples of data in this category are work, education, health and household composition.
- A landmark event. Examples are public and personal.

A.2.1.1. Advantages of Using LHC Technique

- Improve the quality of the retrospective: The use of the LHC concept ensures that events relating to an individual is stored and gathered; therefore the quality of the retrospective data gathered can be improved as the respondent is able to relate both with the timings of the events visually and mentally. This respondent can also use the full patterns of their life events which were recorded to remember the timings of past events accurately (Perl 2002).
- The nature of LHC enables gaps in the data to be detected thereby Improving completeness of reports.
- The calendar format and the visual cues help interviewers collect complex life history data by structuring questions within timing and sequencing framework, while flexible recording techniques allow interviewers to adjust their questions in accordance with respondents' ability to recall events (Perl 2002).

2.6.4. Method Proposed by Nigam and Hurst

In order to determine or detect expressions behind a topic or comment made by an individual, Nigam and Hurst (2004) proposed a method that involves two distinct components. These are:

- Shallow Natural Language Processing Polar Language Extraction System (SNLPPLES).
- Machine Learning Based Topic Classifier (MLBTC).

These components work hand in hand to aid in making a simple and accurate assumption which states that if a topical sentence contains polar language, the system predicts that the polar language is reflective of the topic, and not some other subject matter (Nigam & Hurst, 2004). In this method,

the polar language is first identified in the individual sentences by the system, the machine learning based Topic classifier is then trained with machine learning techniques using a collection of documents which relate the topic, the document classifier is then adapted into a high precision/low recall sentence classifier.

During the analysis, if the classifier judges the document as irrelevant, then all the sentences in the document are assumed to be irrelevant as well; but if the classifier judges the document as topic-relevant within the topic specified, then the sentences are assumed to be relevant, further examinations being then carried out on the entire sentence in the document. Nigam and Hurst (2004) went further to use the Bayesian statistics to form an aggregate authorial opinion metric. This metric was used to propagate the uncertainties which could be introduced by polarity in other to statistically facilitate valid comparisons of the various opinions across multiple topics. The Nigam and Hurst theory was used by Mishne and Glance (2005) to predict movie sales from blogger sentiment.

The concept proposed by Mishne and Glance (2005) is judged unsuitable for the purposes of this research, because it relies on file based content, and that does not apply readily to data from social media networks.

The concepts that have been reviewed thus far would be applicable if the goals of this research were non-exploratory. In other words, if the goals of the research were to generate a consumer confidence index based on a single source of data the above approach would add value. In spite of this, knowing that there are a variety of concepts that could be used to extract information. In the next section, the main tool chosen for this study to extract data from social media networks will be reviewed.

2.7. Data mining tool

2.7.1. About NetBase®

NetBase is capable of extracting data which contains both an aggregative indication of occurrence (across random demographics) and raw natural language text posted on different social media networks. NetBase implements the concept of Natural Language Processing (NLP) and it uses it

to determine the bias of a comment, opinion and even emotions expressed on social media networks (NetBase, 2014).

The NLP is very effective as it does not count the words or analyze word for word like a conventional method such as LIWC. Instead, NLP uses the entire sentence, then the patterns of the grammar/sentence are analyzed before being organized (GNIP, 2010). It uses two specific steps when analyzing the data:

- *Parsing:* This is the first phase in the analyzing stage. In this stage the NetBase NLP engine parses each of the sentences gathered from the social media. It then identifies the various subjects, objects, verbs, adjectives and other patterns in the sentences at the analyst‘s discretion, thus enabling a deep and accurate understanding of the comment to be extracted (GNIP, 2010).
- *Normalization:* This involves defining a standard or an algorithm for how the sentence would be analyzed. At this phase the NLP engine defines the appropriate algorithm for matching keywords thereby preventing inaccuracy (GNIP, 2010).

2.7.2. Advantages of Using NetBase®

- *Accuracy***:** NetBase uses a sophisticated Natural Language Processing (NLP) engine for gathering and analyzing the data (Netbase, 2012). This engine ensures accuracy in the data collected. According to a report, NetBase is nearly 90% accurate (GNIP, 2010).
- *Speed:* NetBase guarantee real-time data and analyzes as it is carried out on social media around the world.
- *Ease of Use:* It is very easy to use. No required experience is expected to be able to read the data as the software performs the gathering and analyzes. It also groups the comments as positive, negative and neutral thereby allowing the analyst to be able to measure the impact of their brand or organization across the various social channels.

2.7.3. Limitations of NetBase®

- *License is Expensive:* The license key for the software is quite expensive and may not favor small and medium enterprises. Only the large organizations are able to purchase this license as they feel they are at greater loss and they should be aware of what their consumers think about their brands or products.

- *Extracting and Filtering of Data is left at the analyst's discretion*: Not all the data is extracted and filtered. The analyst has the power to determine what should be filtered and extracted, and therefore there is a possibility that they can be biased on their decisions.

There are other data mining tools that implement a concept of data mining mentioned or a derivative of one. The limitation of some of these is that they are specialized and optimally designed for applications with a narrower scope compared to this research. As a result, NetBase® is preferred in the present research because its designed scope of utilization is broad as it implements NLP.

2.8. Chapter Summary

Some awareness of the economic indicator of Consumer Confidence indices has been attained, leading the way to explore how it has been measured historically and how the methods of measurement have evolved up until now.

In an age of social media networks, an opportunity exists to explore this new data source whose content is mainly text based but increasingly encompassing other forms of conversations. This is also the objective of this research—to explore whether this data source is sufficient to synthesize information from a diverse dataset of uncommon language rules, syntax, semantics and structure.

The review of the literature in the field of data mining has shown there is a strong interest in the application of social media networks and their data for analysis of real world problems. As such, in order to harness this data source, concepts such as NLP that have been in the fore of research interest since the 1950s were reviewed and an introduction given to other data mining concepts.

Other advances that have been made recently are in the area of interpretation and association of the content of social media networks with the intentions of the participants. An organization such as NetBase has done well in their application of the concept of Natural Language Processing for information retrieval. By choice it is their user interface tool that was used to extract data from social media networks for the purpose of this research study. In the next section, the methodological implementation of this research is presented.

Chapter 3

Methodology

For this research, social media networks such as Twitter, Facebook, Blogger, Google+ and Instagram constitute the sources of the data collection. Each of these are categorized differently and present a different perspective of human interaction. For example, Facebook allows the posting of a mood status—'what is on your mind'; Twitter encourages the use of 140 letter characters to tweet ones viewpoint on a subject of interest and Blogger allows people to present their viewpoints as a story.

For the current research, a dataset of 500 observation was used. The raw dataset of 500 records was stored alongside the summarization of the data by NetBase® relative to the search term. Each raw dataset contains metadata on the location where it originated from, the source (a website like Facebook, Twitter, and blogs), date, time of the day it was posted and the actual text of a posted communication on any one social media network.

3.1. Social media data collection

Social media networks in which the content is in text format are the sources from which data will be extracted using the NetBase® data collection tool. The tool extracts Tweets, Facebook status, FourSquare check-ins and blogs it associates with particular search terms.

The type of data collection procedure is *search* driven because the primary data collection is automatically executed through electronic algorithms built into the NetBase®. Such algorithms implement the aspects of NLP in order to provide data related to the subject of interest, i.e. the keyword. (The keyword is implicitly related to the subjective disposition of users about their economic situation). The tool selected for this research is NetBase®, already introduced in the review of data mining tools.

Since knowledge of the model assumptions used in NetBase's NLP algorithm is unknown, the research used specific [search] terms to extract data deemed to be associated with *economic sentiment expression* for the purposes of this research. Each search is done for a time period

covering 3 months (Jul. 2013 – Oct. 2013), 6 months (Apr. 2013 – Oct. 2013) and 1year (Oct. 2013 – Oct. 2013).

The so-called "search terms" are the result of extensive research into what constitutes the variables used in the measurement of the consumer confidence index as an economic matter. The choice of these terms stems from their association with other economic terminologies and likelihood of them being used within social discussion circles. A good example is the use of employment, as it relates to the total number of employed persons and the unemployment rate (Roberts & Simon, 2001) .

3.2. Keywords for search

Based on the developed attributes on economic sentiment expression the following terms were used to search NetBase® for data.

1. **Capital Expenditures**— this term is highly used in corporate communication and news media and it has a strong link to investment in physical assets by governments, businesses and individuals with significant asset footprint—machinery or real estate
2. **Capital Investment**— is an investment or expenditure that is expected to generate future economic benefits. It can also be seen as a long-life asset— (Investopedia, 2014).
3. **Consumer Confidence**— this refers to the level of optimism, trust, believe and perception of consumers in the economy. It is an economic indicator that helps measure the mentioned attributes above. According to The Wall Street Journal (2015), it states that the Conference Board had a projection for January 2015, "consumer feel more optimistic about job prospects and future labor markets" – (Madigan, Kathleen ;, 2015)
4. **Employment Condition—**we refer to the level of unemployment and employment rate
5. **Family Prospect—** more likely to be used in informal discussions is the term 'family'. However, there is a reliance on NetBase® to augment both words in order to extend the catchment of inferred circumstances for 'prospect'. When we say 'family prospect' it is a future outlook of a family's economic prospects, whether they afford what they want or not. Example; if a family has an optimistic view or projection on better job, it equals to more shopping and more holidays.

6. **Financial Security—** here we mean shares, bonds and derivatives. All these are resources that can be used to support living standards both now and for future needs. It also means being
7. **Health**— given the ubiquity of expressions of illness or of wellness, there is a direct effect of a person's health situation on their financial cash flow. At a higher level, a general talk of the health of a person, an organization, government or community is also indicative of the economic situation of the subject of interest.
8. **Income Sources**— this can be referred to as various sources of earnings that accrued to consumers.
9. **Liabilities**— represent claims against the asset of an entity. This can also refer to borrowings from a financial institution or person, for the development of one's economic status that is usually payable with a short term or long term plan.
10. **Political Affiliation—** Refers to the relationship that exists between members of a group that seek to achieve common goals, ideas and morals through exercise and gaining of political influence.
11. **Recurrent Expenditure—** this is a term meaning the ongoing expenses made by the Government and other organizations. It could be salaries, wages of workers, scheduled bills such as mortgages and loan payments. Capital assets such as bonds and stocks are not seen as recurrent expenditure.
12. **Retirement**— here we mean people that have retired from active service. After a person's service years with no pension plan or retirement benefit system available, the outlook on that person's economic perspective or opinion will be affected.
13. **Social Status**— refers to various social classes, from elites to masses. A person's social status can be earned through different means: own achievements, inherited positions and primarily by identifying one's character. The primary social character sometimes shapes ones entire life and defines their status in the economy. "One's status tends to vary with social context. For example, the position of a man in his kin group helps determine his position in the larger community"- (Encyclopædia Britannica, 2015).
14. **Tax**— this is a compulsory monetary contribution that is levied by the Government on business profits, employment income, goods and services. The word tax can be ubiquitous. It has a direct and indirect influence on an economy. Example: if the tax percent goes up

in an economy, it can leave the people in a pessimistic state regarding how they view the given economy as a whole.

These fourteen associative terms are deemed as 'dimensions' of consumers' (economic) confidence and they originated from search & extract from social media networks, data that could be used to project an estimation of consumer confidence for the purpose of the study. The selection of these dimensions was based on insights gained by the researcher through a critical review of concepts.

A limitation of this approach is that human communication in text does not always contain key words about the subject of the discussion, its context or the intention of the discussion. Therefore, search results are bound to inaccurately omit posts or blogs or tweets that are classed as having no *correlation* to the context, meaning or intention of the searched word key word.

For example, a status post could indicate that a user has expressed they are *sick/ill/unwell*, which would require an extraction tool to discern whether that user is having difficulty with their bills or job as a result. In order to contain this limitation, multiple search terms were used to extract data from social media networks using NetBase®.

Each search term is objective by selected by its association with the expression of psychological financial sentiment, economic situation news, opinion on economic events and life events that have a bearing on economic circumstance and vice versa.

3.3. Research data analysis method

The extracted data was analyzed to explore its attributes and determine whether it is good enough to measure CCI. The data from SMNs were analyzed in four ways to achieve this goal. The first part of the analysis was qualitative: *Part A*. It focused on the infographic summaries of SMNs data, aggregated by NetBase® automatically. In the second part of the analysis, *Part B*, the numbers NetBase used to arrive at the infographic summaries were analyzed. This was a quantitative analysis that looked into obscure but important ratios. For example taking the volume of the most used word for a particular keyword and dividing it by the number of users of the largest SMN, Facebook, was used to test for representativeness from a quantitative perspective.

In the third part of the analysis, *Part C*, the interest moved to the contribution of SMNs to the data content extracted using NetBase. It was used to identify the SMNs with most contribution and also

to see whether the most popular SMN, Facebook, would also be the largest contributor. Then a fourth type of analysis, *Part D*, focused on the raw data of the actual conversations from SMNs. It looked into the classification of a selection of raw Soundbites to evaluate whether NetBase's classifications are correct. This analysis was also used to identify flaws in the configuration of the data extraction mechanism of NetBase. In particular whether the same Soundbites were collected and so repeated, if they were posted by a single user.

Finally, a statistical analytical tool, KH Coder was used to examine fundamental performance of the NetBase® data mining tool and the NLP model it implements. Especially on its ability to comply with basic language semantic rules, such as the occurrence of words that would normally co-occur in spoken or written format. This fifth analysis was used to gain confidence on the performance of the NetBase® tool at a fundamental level.

The reader is reminded that the premise of this research is fundamentally limited by the model of Natural Language Processing implemented in NetBase®.

3.3.1. KH Coder

This is a software that analyzes text based content quantitatively in service of computational linguistics (Higuchi, 2014). Furthermore it can be utilized for the generation of relationships between words in text which can then be clustered among close or like terms or features (Naoi, et al., 2011). KH coder can be used to perform the following analysis.

1. Compute co-occurrence network of words. That is bringing out words that have the tendency to occur several times and co-relate. It also has the capability to select specific key words to show the co-occurrence of a flow, as shown in Fig. 8.
2. Analyze the correspondence words. This is a statistical technique used to analyze multi-way or two-way tables that contain a measure of correspondence between the rows and columns. It is an exploratory data analysis tool that is used to identify systematic relations between variables in the absence of *a* priori hypothesis about the relationship between such variables. It resembles traditional hypothesis testing. KH Coder also provides options to filter the display of words by Chi-square value, grey-scaled dot, bubble plot (unesco.org, 2008)

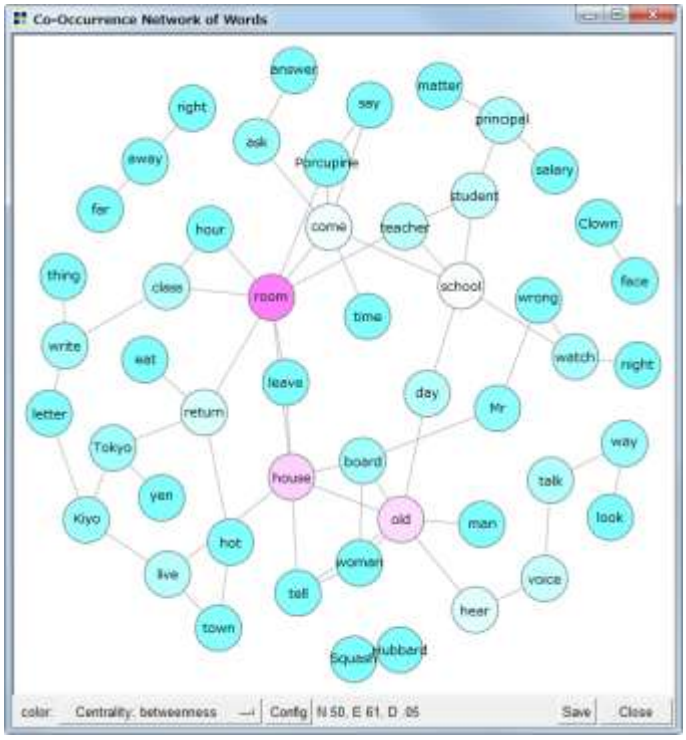


Fig. 7: Illustrating the association of words with other words used with them, an example of co-occurrence network of words. Source: (Higuchi, 2013)

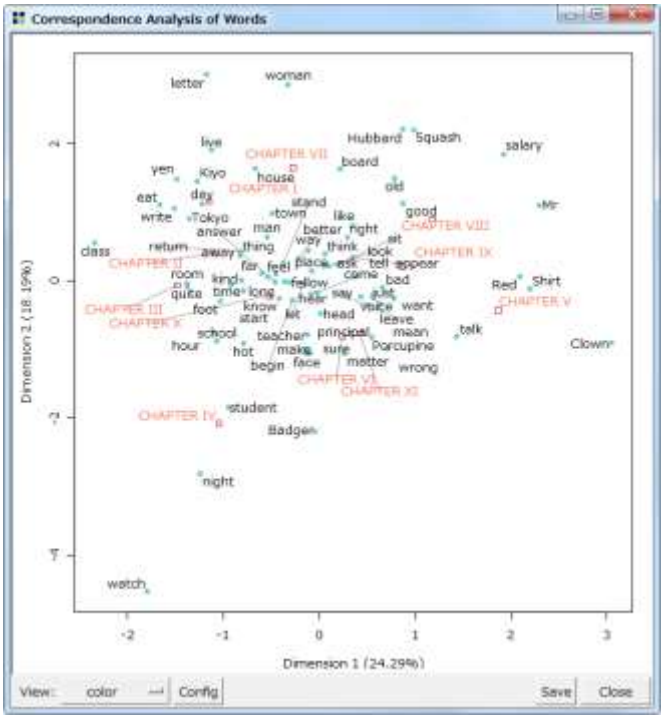


Fig. 8: Illustrating the association of words with correspondence analysis. Source: **(Higuchi, 2013)**

3. Hierarchical cluster analysis uses a variety of levels to group data into a form of cluster tree. The tree is a multilevel hierarchy in which clusters at one level combine to form clusters at the next level. Thus it allows the flexibility to control the number of levels most appropriate for an application.

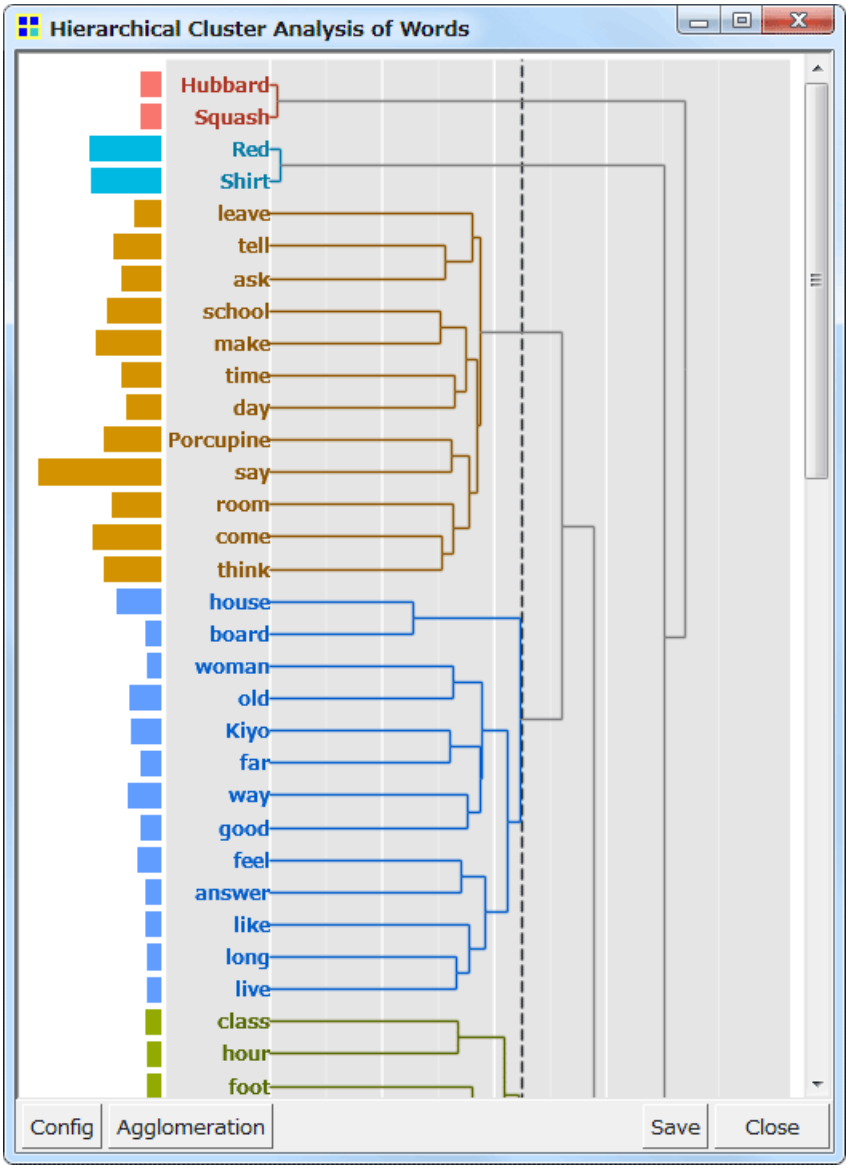


Fig. 9: An example of output from hierarchical analysis using KH Coder. Source: (Higuchi, 2013)

3.3.2. Raw dataset classification criteria

The results or sentiments are classified into four basic categories:

- ***Positive:*** Positive implies that the comment provided by the individual about a specific discussion is 'good'. The positive sentiment is assigned to a comment if that particular comment is factual and the individual agrees with the discussion. For example, if the discussion topic is 'Would you purchase an IPhone 5S Phone'? and the individual's comment to this question is 'Yes, I would definitely buy an IPhone 5S', their response/comment is known as a positive comment.
- ***Negative:*** Negative is the opposite of the positive; it implies that the sentiment behind a comment given in a specific discussion is 'bad' and the individual does not agree with the discussion topic. For example, if the discussion topic is "Would you purchase an IPhone 5S Phone?" and the individual's answer/comment to this question is 'No, I would not definitely buy an IPhone 5S'. Their response would be a negative comment and the sentiment behind the comment is negative.
- ***Mixed:*** This category involves the combination of both positive and negative. In this case the sentiment behind a comment is neither completely positive nor negative, instead there is some form of mixed feelings and condition. For example if the discussion topic is 'Would you purchase an IPhone 5S Phone?' and the individual's comment to this question is 'I might buy an IPhone 5S if I have enough money', that type of comment would be considered as a mixed comment and the sentiment behind the comment is mixed.
- ***Neutral:*** In this case the sentiment behind a comment is neither positive nor negative instead there is no specify answer to the question. For example if a discussion topic is 'Would you purchase an IPhone 5S Phone'? If the individuals comment to this question is 'I am not sure if I will buy an IPhone 5S'. This comment is known as a neutral comment and the sentiment behind the comment is neutral.

In summary, the correct application of NetBase® to extract data from social media networks affects how reliable the analysis of the data will be, where such analysis aims to establish whether the data is good enough for the determination of important information like the economic confidence of consumers, quantified as the CCI. In Chapter 4, the data extracted from NetBase® using the [search] keywords is presented in graphical, qualitative and quantitative forms and analyzed accordingly. Then fundamental attributes of the data tested in the KH Coder tool are presented.

Chapter 4

Results and Analysis

Each keyword that was used as a search term for the period of 3 months, 6 months and 1 year, resulted in three different outputs from NetBase. The first is a graphic summary of the extracted dataset NetBase classified as being associated with the search term. This graphic summary is a word cloud that summarize highlights of the dataset output for a keyword. For instance, the search term 'Capital Investment' for the period of 6 months, gave an associated word cloud that summarized the 'Top Emotions' expressed, the 'Top Terms', the 'Most used hash tags[3]', and the 'Top conversation places'. Please see Fig. 10.

Fig. 10: Word cloud of (a) Top Emotions expressed; (b) Top Terms used and (c) Most used hash tags

3 "The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages" (Twitter, Inc., 2014)

The other output from NetBase is a tabulated summary of quantified measurements of the various insights given in the word clouds. Table 2 shows the 'Top Terms' used alongside the number of times they occurred in social media networks discussions over the period of interest. The top words used to express emotions associated with Capital Investments are also shown in Table 2.

Table 2: A frequency count summarizing the Top terms & Top emotion words related to Capital Expenditure

Terms	Mentions
investors	1207563
Fund	990841
projects	698835
property	681345
stocks	663855
income	636100
industry	588573
Security	538384
Children	462341
return	438008
U.S.	361204
economy	358923
report	348657
assets	324703
money	320901
sector	313297
capital	312537
dollars	296948
firm	281739
opportunities	279078
costs	270713
loans	268051
interest	267291
rate	260067
value	258166
risk	255124
growth	242957
family	235353
bank	228889
laws	225847
member	192769
portfolio	184404
Profit	179081
education	177560
benefits	172617
cash	168055
interest rates	167294
foreign investment	165393
Federal Government	163492
Investment Bank	162351
good investment	148283
data	140299
policies	133835
Income Tax	126991
equity	122809
bonds	117866
China	116726
Return on Investment	112923
hedge fund	108361
development	103798

Emotions	Positive Mentions	Negative Mentions
best	57465	0
good	32875	0
great	26819	0
enjoy	9213	0
prefer	7880	0
interested	6658	0
important	6271	0
like	4598	0
successful	4495	0
happy	3831	0
thank	3819	0
love	3796	0
worth it	3615	0
attractive	3539	0
solid	3491	0
not fail	3314	0
worthwhile	2653	0
look forward to	2361	0
appropriate	2235	0
pleased	2102	0
favorable	1896	0
vital	1634	0
beneficial	1632	0
not force	1591	0
proud	1326	0
blame	0	7778
bad	0	4725
concerned	0	3553
worried	0	3028
accuse	0	2722
hate	0	2515
not want	0	1980
not like	0	1790
fear	0	1782
complain about	0	1400
poor	0	1368
worthless	0	1135
not afford	0	1079
worst	0	1029
worse	0	996
wrong	0	993
fuck	0	986
stuck	0	925
bad bitch	0	887
stupid	0	864
shit	0	862
frustrating	0	822
terrible	0	749
fed up	0	745
tired	0	732

The third output is the raw dataset NetBase® collected from a variety of social media networks and indexed. There are 500 of these for each of the time periods per searched keyword. It contains the actual Facebook comments, Twitter tweets, Blog comments and comments from a variable of other sources. These are called

Sound Bite Text. However, NetBase® also classed these Sound Bite Texts as either being Positive, Negative or Neutral. Other information include the domain name of the social media network where each Sound Bite Text is extracted from, date and time stamps of publication among other details, as shown in Table 3.

For each of the keywords used to extract data from NetBase®, these three sets of outputs are obtained for the 3-month, 6-month and 1-year periods. Given that there are fourteen search terms and nine sets of data outputs per search term, there are total of 126 datasets to analyze qualitatively and quantitatively. Therefore it is important to focus on the objective of the thesis—i.e. to evaluate whether data from social media networks could be used to meet the needs of determining information such as the consumer confidence index. In light of these constraints, seven of the datasets captured for each searched keyword were analyzed instead of all fourteen.

The word cloud summary results shown in Fig. 10 were analyzed qualitatively. Then the tabular results shown in Table 2 were graphically analyzed. The results shown in Table 3 from all fourteen searched keywords were combined and analyzed collectively using statistical methods implemented in the KH Coder tool. These gave different rich perspectives on the attributes of the data from social media networks, their strengths and weaknesses if they were to be used to measure a metric like CCI.

Table 3: The third output from a NetBase search; the indexed dataset extracted from social media networks for the search term of Capital Expenditure

Sequence	Sound Bite Text	Sound Bite Sentime	Likes	Dislikes	Pos Emotions	Neg Emot	Pos Behav	Neg Behar	Title	Source Ty	URL	Domain	Published	Author Ge	Author UI	Author Na	Author Ha	Author Lo	No. of Fol	Klout Scor	Similar
1	Foreign investment in	Positive								Social Net	http://ww	facebook.	2013-10-1		http://ww	Ibiza Soth					No
2	This is in spite of the f	Negative		higher price		blame			Norwegia	Blogs	http://ww	telecompa	2013-10-1						50		No
3	Interesting: Tory Coun	Negative		homelessne		blame				Social Net	http://ww	facebook.	2013-10-1		http://ww	Growth - M					No
4	Have the SNP not mad	Positive	outperform UK							Comment	http://ww	heraldsco	2013-10-1				p4ked				No
5	I did hear (not sure wh	Positive							Re: Movin	Forums	http://bri	britishexp	2013-10-1		http://bri		AspiringC				No
6	Bottom line, my inves	Positive	greater under Oba							Comment	http://ww	nationalre	2013-10-1				agwtub				No
7	I think that makes it qi	Positive							Changing	Blogs	http://ww	interest.c	2013-10-1			Gareth Va		New Zeala			No
8	More than 1,200 peop	Positive							Migrants a	Blogs	http://ww	myheadli	2013-10-1			BBC News					No
9	&& det lil investment	Positive	smart choice							Microblog	http://twi	twitter.co	2013-10-1		http://twi	†_Prince B	_TapOut_	Tennesse	1118	35	No
10	House prices in Londo	Negative								Comment	http://ww	frenchyun	2013-10-1				ad0n91				No
11	House prices in Londo	Negative							Flipping H	Comment	http://ww	frenchyun	2013-10-1		https://pl	Ann-Marie					Yes
12	Now they're trying to	Negative							'Snowden	Blogs	http://ww	dw.de	2013-10-0					Germany	6988		No
13	Hasan Malek, Domesti	Positive							HARAKAH	Blogs	http://haf	hafifi-hafi	2013-10-0		http://ww	Admin					No
14	Hasan Malek, Domesti	Positive							BERITA TE	Social Net	http://ww	facebook.	2013-10-0			Harakahda					Yes
15	"So, he (Xi) will consic	Positive							見高拜,易	Forums	http://cfo	cforum.ca	2013-10-0		(http://cf		jvcpcv55	Malaysia			No
16	"So, he (Xi) will consic	Positive							When a sr	Blogs	http://life	lifeofaanr	2013-10-0	Female	http://ww	Annie					Yes
17	"So, he (Xi) will consic	Positive								Blogs	http://ww	thebome	2013-10-0						341		Yes
18	"So, he (Xi) will consic	Positive							DR MAHA	Blogs	http://kw	kwongwal	2013-10-0					Malaysia	55		Yes
19	So, he (Xi) will conside	Positive								Social Net	http://ww	facebook.	2013-10-0		http://ww	Tima Inter					Yes
20	"So, he (Xi) will consic	Positive							Dr M sugg	Blogs	http://allr	allnews.rs	2013-10-0			editor@th			587		Yes
21	"So, he (Xi) will consic	Positive							Dr M sugg	Blogs	http://allr	allnews.rs	2013-10-0			editor@th			587		Yes
22	This is the problem th	Negative		create prob						Comment	http://mo	money.cn	2013-10-0				02w84		150747		No
23	RT @_G_SWAGG: I'm r	Positive								Microblog	http://twi	twitter.co	2013-10-0		http://twi	E.K.	amorttay		917	40	No
24	less than half the price	Positive								Comment	http://ww	motherjor	2013-10-0				zhsci		18657		No
25	Business investments	Positive								Comment	http://ww	thebome	2013-10-0				ad04w2		341		No
26	We have people feeli	Positive								Social Net	http://ww	facebook.	2013-10-0	Male	http://ww	Curt Dooli					No
27	RT @martabausells: "I	Mixed		addiction, t	addiction					Microblog	http://twi	twitter.co	2013-10-0		http://twi	Jawwad Fa	rebootduc	Pakistan	596	49	No
28	Recipient France prais	Negative							Luxury TV	Blogs	http://ww	dw.de	2013-10-0					Germany	6988		Yes
29	EU top court rules agai	Negative							Markets s	Blogs	http://ww	dw.de	2013-10-0					Germany	6988		Yes
30	Barroso sees EU meml	Negative							German jc	Blogs	http://ww	dw.de	2013-10-0					Germany	6988		Yes

4.1. Assessing the robustness of social media networks data

The analysis is divided into three sections based on the type of output obtained from NetBase. The first section is a comparative analysis of the infographic word clouds output by NetBase: Part A. The aim is to examine how these summaries compare to the mainstream economic discussions during each period of time. The next part of the analysis follows from the first part, and entails the visualization of the data shown in Table 2. The aim of the visualization is to explore the aggregate summaries made by NetBase: Part B. This happens to be the numeric values behind the infographic data from NetBase and shown in Fig. 10. It gives a quantitative perspective of indicators such as—how many times a vocabulary was used across social media networks over each period of time and whether there are notable changes in vocabulary used. In Part C, the largest sources of SMN data were identified, each with its own contribution of content for each keyword discussed. Then in Part D, the raw data indexed by NetBase was analyzed. This analysis was used to assess different aspects of how NetBase as a tool performed. For instance how it classed different Soundbites—*positive*, *negative* or *neutral*—and whether these are deemed the correct classification. This analysis also looked at the interval between SMNs data extractions and the number of Soundbites that were repeated, even when they are not from the same source.

In the last part of this analysis, statistical analytical methods were used to explore fundamental attributes of the raw data from NetBase. This was to check that the classification performed by NetBase meets basic language semantic rules. For example whether two words that are used together, occurred together in a Soundbite. This was achieved using the KH Coder tool in section 4.2.

4.1.1. Part A: Qualitative analysis based on Infographics

The infographic summary from NetBase is collated for seven of the fourteen keywords. Each one has ten word cloud summaries for the 'Most used hashtags', 'Top Attributes', 'Top Behaviors', 'Top Brands with the category', 'Top Emotions', 'Top Likes', 'Top Terms', 'Top Conversation Places', proportion of 'Gender' contributions and 'Top Dislikes.'

For the purposes of this research each of these is evaluated based on two criteria to streamline the scope of analysis and keep it focused. These are (a) whether the category gives an insight into the vocabulary that is directly associated with the context of the search term and (b) whether it has a contextual relevance to the expression of human sentiment as it relates to their economic situation. This is because the notion of consumers' economic confidence is a sentimental one.

Hashtags have become a dominant feature of written human expressions in this era of social media networks. It is used to associate expressions to any subject of interest, popular or unpopular. Therefore it is one of the categories that would be analyzed because it meets both criteria. The other categories are 'Top emotions' and 'Top Terms' which meet the requirements of both criteria.

From the literature review it is known that human conversations rarely reference the *formal name* of a subject of interest being discussed. For instance, people express their opinion on a government policy by making cynical statements that do not contain a word from the policy document. Or better still they use swear words or slurs to express how they feel.

This is a constraint for this research because terms such as: consumer confidence, political affiliation, employment condition or social status, are minimally used terms in informal conversations. Therefore, to mitigate this limitation, two keywords were chosen—one is used frequently in informal conversations and the other is used frequently in formal conversations. This choice was made to facilitate a richer analysis.

Tax, a frequently used word (in both informal and formal conversations) is one of the keywords of choice. The other keyword is *Capital expenditure*, a frequently used word in formal conversations. It was hoped that analyzing the data generated by these keywords would throw some light on the strengths and weaknesses of data from social media networks. The other keywords were analyzed collectively because they fell into one of either category of being formal or informal. Therefore they will share similar attributes to either one or the other of the two keywords selected for analysis.

4.1.1.1. Tax

From the literal definition of tax[4], everyone has the obligation to pay some form of tax. The word/term is also used in formal and informal conversations and it attracts direct feedback of opinion, especially in the event of a change of government tax policies. Various forms of taxes include the Value Added Tax which impacts on the cost of living via the cost of goods & services; corporation tax—on the profits of companies; sales tax—usually added to the basic price of a commodity; import-export duty tariff—on goods brought into or taken out of the country or taxation—in the form of the PAYE that is deducted from peoples wages when they are paid (The New York State Department of Taxation and Finance, 2014) (Investopedia LLC, 2015).

The qualitative outputs for the tax keyword in Fig. 11 identifies words that are related to 'tax' and were used most often. In the category of 'Top Terms' and 'Most used hashtags', these are orange in color, while in the category of 'Top emotions expressed' these are the green words.

4 A compulsory contribution to state revenue, levied by the government on workers' income and business profits, or added to the cost of some goods, services and transactions (Oxford Dictionaries, 2015)

Their font size represents their relative frequency of use. A visual inspection reveals that the top terms associated with the Tax keyword are vocabulary related to our monetary economic system, its institutions and activities pertinent to their operational existence.

Starting with the key economic player i.e. the Government (the top term), to activities that influence the economic situation of everyone at a personal level; e.g. commercial or public sector 'projects' that generate cash flow across the economy, 'investment' in infrastructure by Government , 'funding' of a variety of programs as part of government policy, 'income' and 'pension plans'. There are also entities that influence the state of the economy through the use of economic levers. These are the 'banks' and their 'interest rates', political parties—'Republicans' or 'Democrats' and how their leadership of government could mean a change in 'government funding' policy, exchange rate of the 'dollars', the 'costs' of running government departments or the plan for 'debt' management.

So far as a government is concerned, taxation is a fundamental mechanism to incentivize or deter economic actions of any kind. This is because the government's policy programs, departments, projects and labor must be financed. What this means for other economic players—from corporations to small businesses and individuals- is that they are obliged to fund the government through tax. Given that tax is one income channel used by governments to finance their recurrent expenditure, like debts, salaries, pensions and capital expenditure, tax affects the economic outlook of everyone in an economy.

Agreeably, these 'Top terms' associated with Tax as an aspect of a consumer's economic perspective do include popular terms used in mainstream economic discussions. The terms are contextually relevant, however they only give an indication of the popular terms related to Tax, not how consumers feel about 'tax', which in turn will affect how they feel about their economic situation.

Even if the "Most used hashtags" were visually inspected for a hint of the general mood at the time, it would be found lacking. This is due to the fact that the most used hashtags—'#funding', '#savings', or '#cash' give a non-directional perspective relative to human situations. In that they are identified implies the topics of most interest in that period, but give no clue as to whether people had more funds available at their disposal or whether there were more talks of need for funding.

Although data from social media networks easily provides us with insights into what is being said by those who participate in it, it may not give us the context or intentions of their interactions. This is attributable to the two forms of emotions which humans experience—the *incidental*[5] and *integral*[6] forms of emotions, and how they influence our judgment of any situation, including our own (Daas P.J.H and Puts M.J.H - European Central Bank, 2014). This is where new tools like NetBase® are of crucial help.

---

5 & 6 "Incidental emotions are the emotions we carry with us to a decision that have nothing to do with the decision." "…integral emotions are emotions that are caused by the decision itself." (Karen Christensen - Forbes India, 2014)

Fig. 11: Qualitative summary attributes of data associated with Tax keyword

Incidental and integral forms of emotions are responsible for the lag in the expression by Dutch users of Facebook and Twitter SMNs (Daas and Puts 2014). In the same work, they refer to an earlier work where 50% of the data acquired from a major social network platform, Twitter, is deemed "pointless babble." This prospectively forewarns that some of the data from SMNs is noise, and without contextual relevance to the subject of interest (Daas P.J.H and Puts M.J.H - European Central Bank, 2014).

These findings are further buttressed by a Facebook research about *emotional contagion* (Kramer, et al., 2014). It found that people tend to post positive or negative looking conversations depending on the overall "feel" of what their social media friends express and the most recent news feed from their Facebook network on their Home page.

Nevertheless, it is inadmissible to dismiss half of the dataset we have as irrelevant because they are associated with economic discussions we all participate in, one way or another. Thus, the missing knowledge here is how to tell the level of accuracy of the model that was used to arrive at this 50% noise claim.

With this awareness, we now look to know how consumers feel about 'tax'. We look to the 'Top emotions expressed' for each period as shown in Fig. 11. Not only are the words 'great', 'enjoy', 'good', 'love', 'best', amazing' striking as being positive expressions, overall the high proportion of greens (positives) compared to red (negatives), imply a positive economic outlook by the largest portion of the sample.

One could postulate that the positive expressions could be induced by some ongoing incentives highlighted in the 'Top terms.' For example, financial 'Benefits' in form of tax credits given to individuals could induce the expression of 'good' times, a time to 'enjoy' favorable 'investment' plans by private corporations.

In as much as some causal association could be postulated to claim that the Top emotions expressed tell us that consumers are feeling happier or more confident— economically based on the Top terms used and the events surrounding them— the data does not actually support such postulate. Instead, the data gives a linear association between the most expressed emotions found in social media networks interactions that are associated with Tax, the searched Keyword.

On close observation, these summary infographics seem to originate mainly from North America given the type of terms seen. For example there is the #tcot and #obamacare hashtags; then 'Republicans', 'Democrats', 'Bush' and 'Senate' Top terms synonymous with North America, the U.S.A in particular. Henceforth, it will be assumed that NetBase captured data from the North American region.

Let us look at the 'Most used hashtags' once again. Ignoring observed duplications, hashtags such as #funding, #cash, #savings and #invest do point at a fact that there was a buzz on social media networks about the economy and Tax. The implicit requirement to "trust" the classification done by NetBase based on fundamental Natural Language Processing principles is tempting though attributing relationship is always dangerous. It could be that the data classified was either contextually relevant to the searched keyword, or contextually relevant to a higher level of knowledge, in this case the [sentimental] economic confidence of consumers who are also users of SMNs. An external validation of the 'mood' of consumer sentiments around each period of interest is one way to evaluate these limitations.

This could be achieved by using published CCIs between July and October 2013. However, this research does not measure consumer confidence, and therefore such external validation of measurements of Consumer Confidence is excluded from this analysis. Nonetheless, this need for external validation points at a weakness of lack of causal association beyond the direct association with a keyword used to search.

In summary, unlike CCSs where the context for questions posed to correspondents has a set and clear scope, interactions on social media networks are diverse, varied and random in nature. Although corporate institutions and businesses tend to post communications that are directly linked to a subject of interest, this is not the case for individuals. People tend to post in *response* to an event or topic of interest being discussed. They do not necessarily state how such events or topics affect them personally—at that moment or how it will in the future.

Based on the qualitative analysis of the Tax keyword above, an attribute that cannot be taken for granted with data from social media networks is that the use of the *right* terms will yield the extraction of the *right* data, where the challenge is for the user to develop a model of association of the search terms with the knowledge they seek to derive from the data.

From the perspective of the keyword "tax" and its effect on the economic confidence of consumers, two observations have been identified from the analysis of Fig. 11: (i) that the data from social media networks is to a degree contextually associated with the searched keyword but (ii) give no indication of causality between the data and the higher order knowledge for which it is needed. This poses a fundamental question of how reliable social media networks data could be used beyond the linear association of contextual vocabulary.

This suggest a review of the results of analysis for the other keyword that is used extensively in formal discussions—*Capital Expenditure.*

4.1.1.2. Capital Expenditure

This keyword, unlike the Tax keyword, has a longer term footprint and induces a different kind of consumer reaction. It is described in Investopedia as:

> *"Funds used by a company to acquire or upgrade physical assets such as property, industrial buildings or equipment. This type of outlay is made by companies to maintain or increase the scope of their operations. These expenditures can include everything from repairing a roof to building a brand new factory"* (Investopedia, LLC, 2015).

Capital expenditure spans beyond the precinct of private businesses or organizations. It is also undertaken by governments, international financial institutions and non-governmental institutions amongst others.

A visual analysis of the data yields these insights. The key insight that when it comes to the Capital Expenditure, government is the key player (in the 'Top term'). It can also be seen that factors that influence capital expenditure investments are also contained in the social media network interactions.

In the 3-month, 6-month and 12-month periods, such factors identified are 'fuel prices', 'oil price', 'subsidy', 'inflation', 'house prices' or the 'dollar'—all of which are referenced when governments or organizations announce strategic, long term investment decisions. For instance the drop in oil prices to as low as $50/barrel has led major oil & gas companies to even go as far foregoing the development of lucrative facilities recently. Shell also recently scrapped its $6.5 Billion USD project in Qatar as a result (Sergie, M.A.; Kayakiran, F.; Bloomberg, 2015). These actions have ramifications for consumers because cash flow generated by new projects would have reached host communities and create more jobs and new income streams.

A glance at the Top emotions expressed in Fig. 12 gives the impression that things are 'cheaper'; that there is a 'low gold price'. Also it can be seen that the underlying theme here is the costs of energy being an influence on capital expenditure decisions. A visual evaluation of the Top Emotions expressed shows negative sentiment (in red), and is collectively higher than positive ones (green)—heralded by the high occurrence of the word 'heating cost'.

Assuming there is contextual association, the 'top emotions' expressed hint that consumers are unhappy about the cost of energy and this is validated by the 'top terms' shown in Fig. 12. Furthermore, the Most Used hashtags point at topics of highest interest, which includes government operations (#shutdown, #obamacare), energy (#gas, #syria) and elections (#ronpaul, U.S presidential candidate 2012).

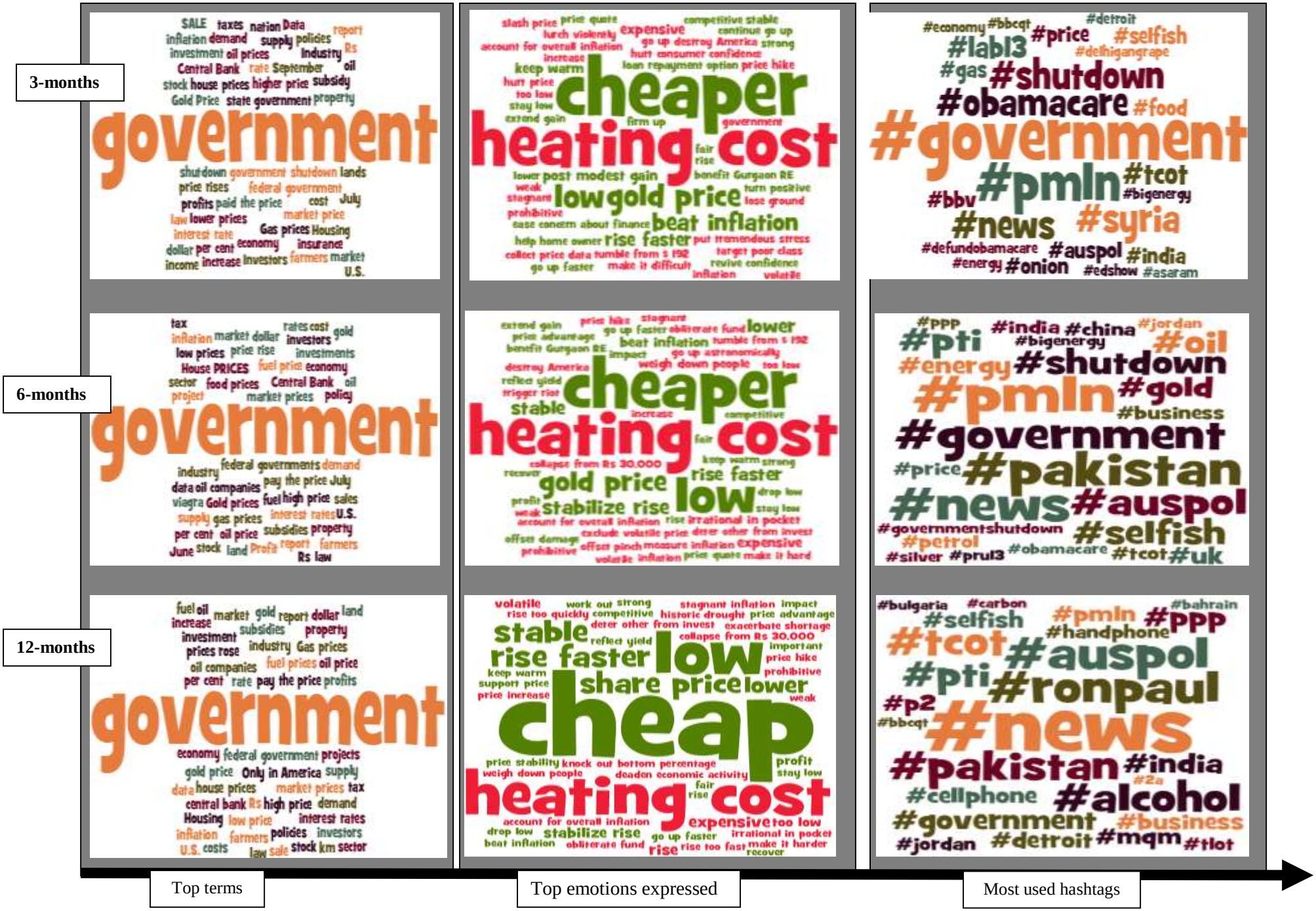


Fig. 12: Qualitative summary attributes of data associated with the Capital Expenditure keyword

Figure 12 could be interpreted to mean that energy prices are unfavorable and consumers are unhappy about it. On the other hand the majority of the topics of popular interest dwell on subjects other than on the Top term ‘government’. Based on this lack of clarity, these hidden links of causality or lack of it, are an inevitable challenge for anyone who intends to use this sort of data to derive knowledge of a higher order.

Without a causal hint to tell whether Top emotions expressed relate to how consumers feel about their economic situation, social media networks data of an unaligned context could be used. In other words without knowledge of the context of the interactions from which the data originated, the implicit bias of the data might be counter useful for the purposes for which it was extracted in the first place.

Irrespective of the bias of the keywords used to extract data from social media networks, the right keywords will generate any sets of data that are relevant. The onus is on the researcher to develop a systemic association between the keywords and the knowledge being sought, with an implicit assumption that contextual relevance is achieved to some degree through the model of data mining implemented.

If contextual relevance is achieved by up to 50%, application and research shows that the data from social media networks is *'good enough'* to support the synthesis of a higher order level of knowledge—albeit from a qualitative frame of reference. This was demonstrated in the work of Daas and Puts (2014) for the European Central Bank. They claim they were able to measure the economic confidence of consumers based on data from Facebook and Twitter, with a correlation of $r = 0.9$.

In 2013, the Netherlands was identified as having the most users on social media networks in Europe. In relative terms, about 65% of its population use social media networks (Office of National Statistics, 2013) (**Error! Reference source not found.**). With such high proportion of SMN users, if it is assumed that at least 80% of the 65% actively use social media, then it is reasonable to assume that online conversations originate from about 50% of the Dutch population. Again, this is relative and not supported by actual data.

In the situation described here, the other half of the population’s economic prospects are unknown to the researchers. This implies that a categorical statement of accurately measuring consumer confidence based on social media networks alone cannot be an end by itself. It is rather inconclusive and immaturely dismissive about the fact that 50% of the population’s views are unknown, and cannot be said to be positive or negative. The insights gained from analysis of Figures 13 and 14 in subsections 4.1.1.1 and 4.1.1.2, are deducible from the rest of the data gathered on the other five keywords shown in Fig. 13, Fig. 14, Fig. 15, Fig. 16 and Fig. 17.

In the same light, the infographic outputs for Employment conditions, Health, Consumer confidence, Political affiliation & Retirement shown in Figures 13 through 17, show the similar flaws highlighted above.

In particular, it can be seen that the ‘Most used hashtags’ do not reflect in the strongest sense, and that the most used hashtag is related to the searched keyword. For instance, the infographic of Health keyword shown in Fig. 16 illustrates this. Other infographics however give an assuring alignment with the keyword.

Holistically, the data from social media networks is varied as each searched keyword yields a different perspective of vocabulary that is different from the others. This is also the case when one looks across each time period, beyond the most frequent top term, top emotion expressed and most used hashtag.

There is an opportunity to advance models that would provide a second (and higher) order association of these data with higher levels of informed knowledge.

The analysis undertaken so far has already identified one major challenge in the area of data mining identified in the literature review: that the accuracy of the data models to associate words to the right *context* and to capture the *true intentions* of the interactions that contain such words on social media networks represent the key determinants to the usefulness of the data.

In summary, the qualitative analysis of data from social media networks hints that this data is 'good enough' based on real-time relevance. However this will only hold if the fundamental assumption of contextual relevance holds.

The next part of the analysis looks into the quantitative aspect of these summarized infographic outputs from NetBase®. The aim is to assess whether each category such as the ‘top terms’ NetBase® has identified is supported by the actual number of times such ‘top terms’ occurred. Other attributes of the data that are looked into includes the volume of online interactions, e.g. tweets, Facebook messages/comments/statuses and the trend of volume over each period of interest.

Fig. 13: Qualitative summary attributes of data associated with the Employment Condition keyword

Fig. 14: Qualitative summary attributes of data associated with the Health keyword

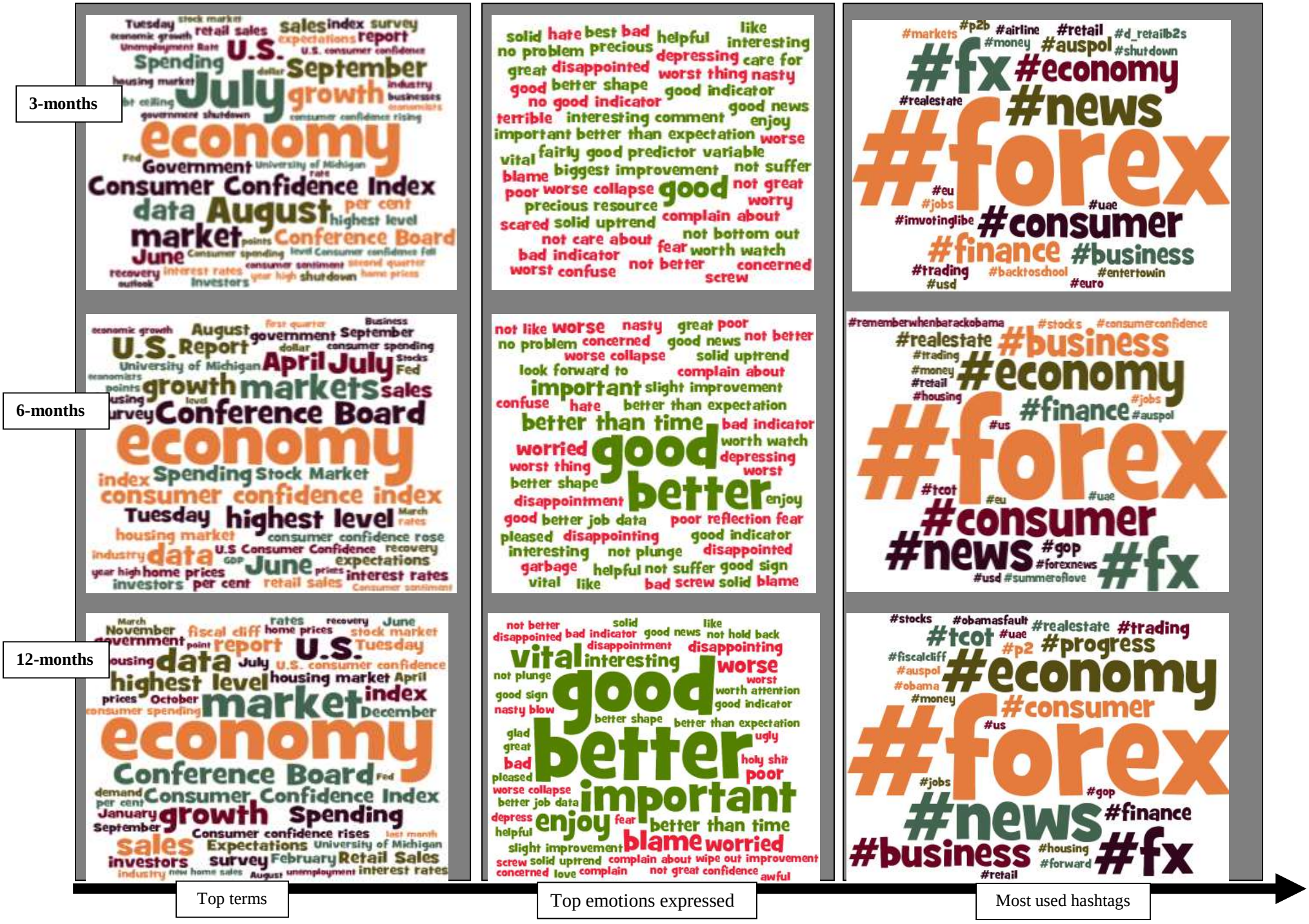


Fig. 15: Qualitative summary attributes of data associated with the Consumer Confidence keyword

Fig. 16: Qualitative summary attributes of data associated with the Political Affiliation keyword

Fig. 17: Qualitative summary attributes of data associated with the Retirement keyword

4.1.2. Part B: Quantitative analysis of aggregated summaries

Tax

The second output from NetBase® for each of the seven keywords selected for analysis is shown in Table 2 (p.42). Analyzing the numeric values yields no hidden trends. As a result, each table obtained for the three subcategories under review is graphed to visualize the data. It is believed that doing so will enable trends in the data to be observed easily.

The Top terms associated with the Tax keyword are shown Fig. 18 (a) 3 months, (b) 6 months and (c) 12 months. The reference point here are the first five terms in each period of time and how often they occurred in the data found on social media networks. Also, the bottom five "top terms" are analyzed to see whether there any observed movements, such as a change in position and magnitude of occurrence. In a similar fashion, the top emotions and the most used hashtags are analyzed.

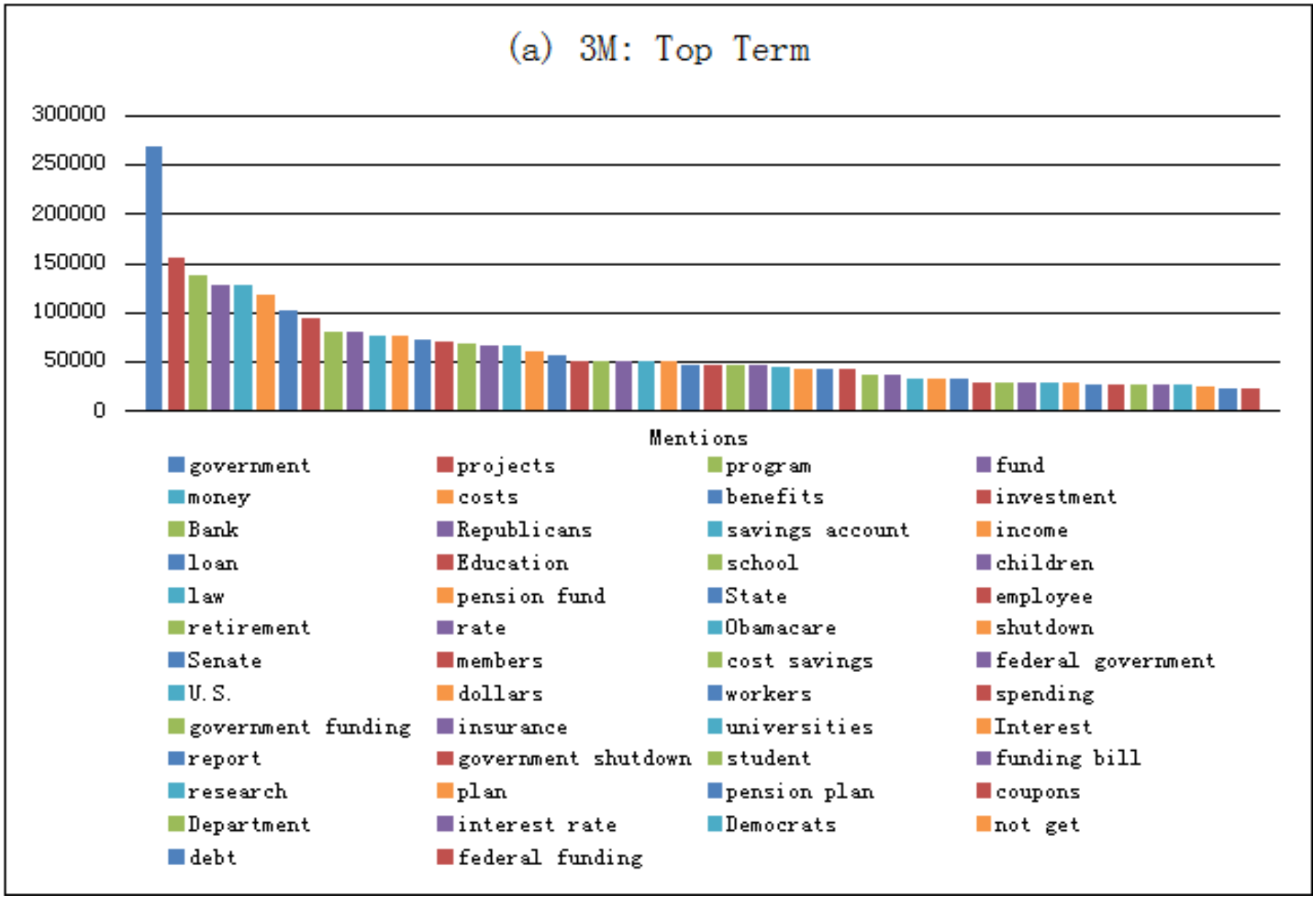
Similar to the qualitative results analyzed in section 4.1.1, in the top terms, 'government', 'program', 'money', 'benefits' and 'bank' are the top five most used words in the three month period. In the six month period, 'costs' and 'children' displaces 'bank' and 'money' as the most occurring terms in online conversations. In the 12 month period, 'fund' which could be qualified as 'money' surfaces, but would be disregarded. This means that 'education' was one of the most talked about topics.

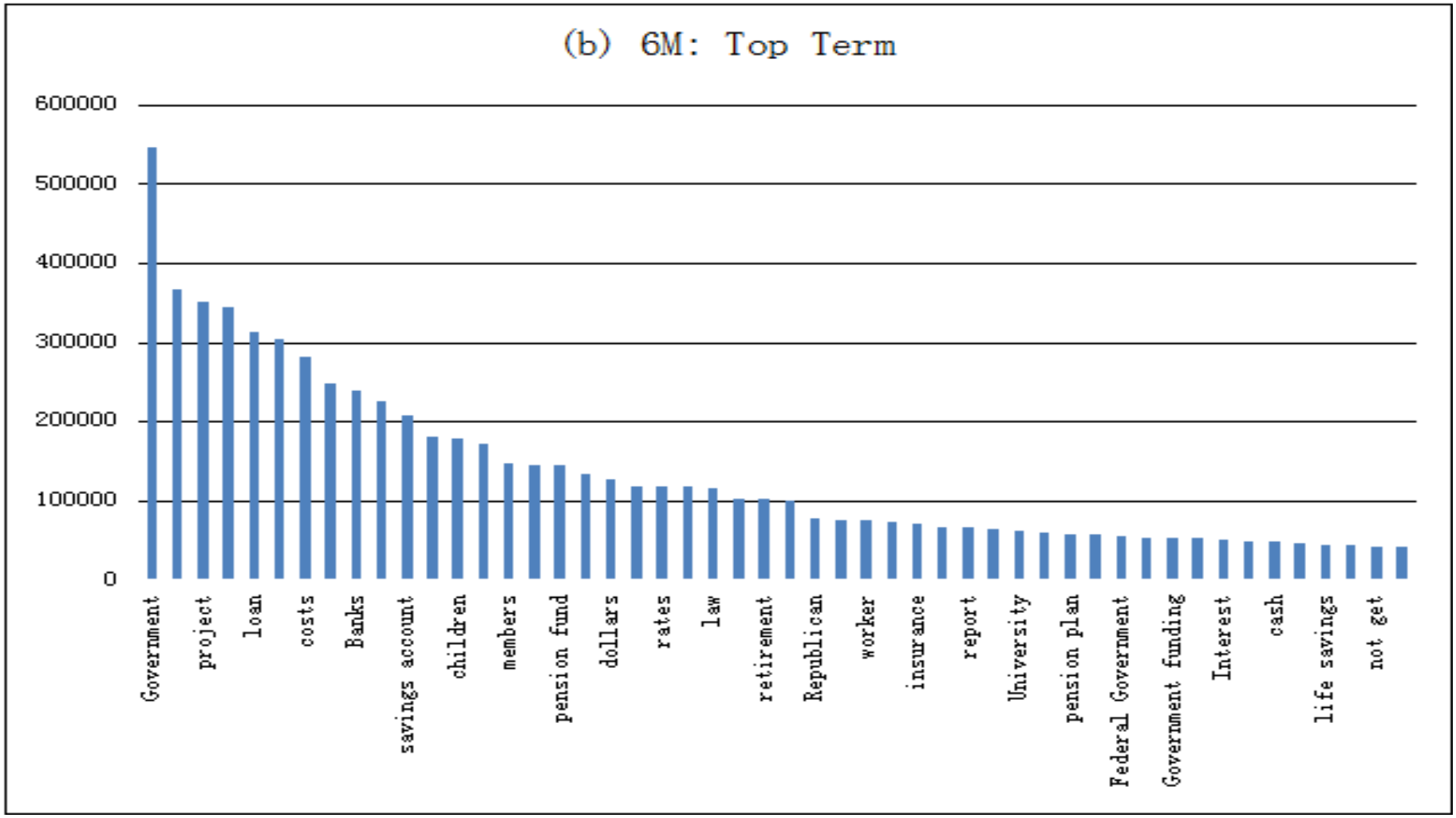
On the other hand the least talked about topics in descending order, in the three month period are 'pension fund', 'children', 'income' and 'republicans'. In the six month period, in descending order, these are; 'not get', 'student', 'interest', 'debt', and 'pension plan'. While 'best vacation ever', 'federal funding', 'plan', 'U.S' and 'pension plans' are the least five mentioned topics in the 12 month period.

From the data, the particular area of interest seems to be that of child benefits. This summary is drawn from the observation that if we look at the three month period and then the six month period, children which was a least mentioned subject became a top five subject of discussion. This observation also holds for education, which became a top five topic if a year's worth of data is considered.

On another front, the five least discussed topics show no associable shifts that could hint at what may be happening, except for the fact that pension surfaces in the least talked about terms and no topic has moved from being highly talked about to being a least talked about topic. If this observation of the least talked about terms holds in the analysis of the top emotions and most used hashtags, only the five top terms will be the focus of the analysis.

Tax: Top terms




(b) 6M: Top Term
600000
500000
400000
300000
200000
100000
0
Government
project
loan
costs
Banks
savings account
children
members
pension fund
dollars
rates
law
retirement
Republican
worker
insurance
report
University
pension plan
Federal Government
Government funding
Interest
cash
life savings
not get

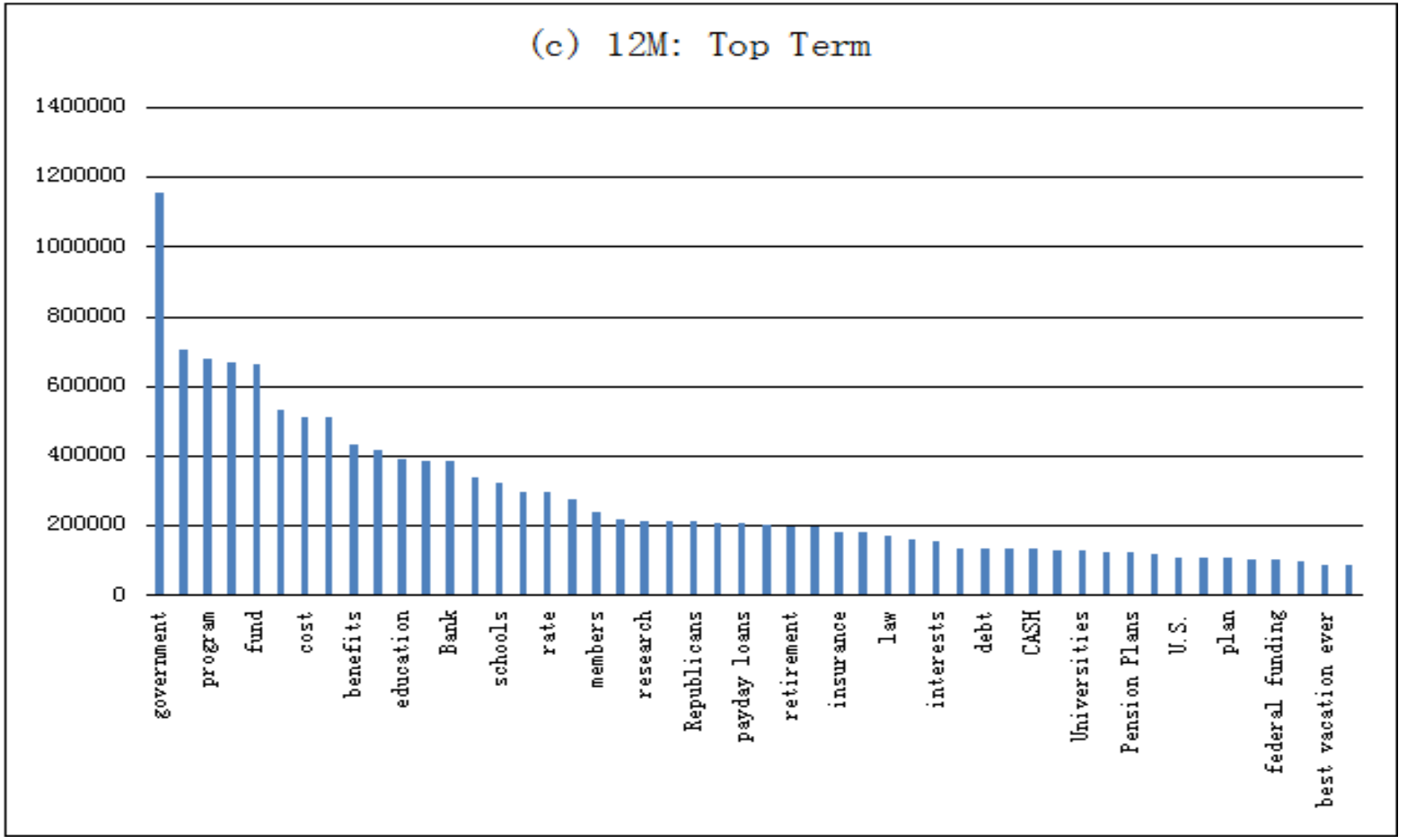


Fig. 18: Graphed Top term representation of the quantified attributes of data from social media networks, associated with Tax

During 2012 – 2013, the top term ‘government’ was mentioned more than 1 million times, while the other four top terms were discussed in about 600,000 conversations during the same period. These numbers are huge, but a relative comparison to the number of active social media network users on Facebook alone, brings to mind the reality of modeling errors. This is because about 1 billion people actively use Facebook each month (The Next Web, Inc., 2013). As such one is right to anticipate that the number of times ‘government’ is mentioned in one year is about 0.12 percent of the active monthly users.

The observation above suggests that, of the 1billion monthly users who use Facebook, that at least the proportion of comments, tweets, messages or blogs that had conversations relevant to the ‘government’ is 1%. Perhaps until a measure of inaccuracy of the NLP data mining model used in NetBase® can be quantified, this estimation of 1% contains a degree of uncertainty.

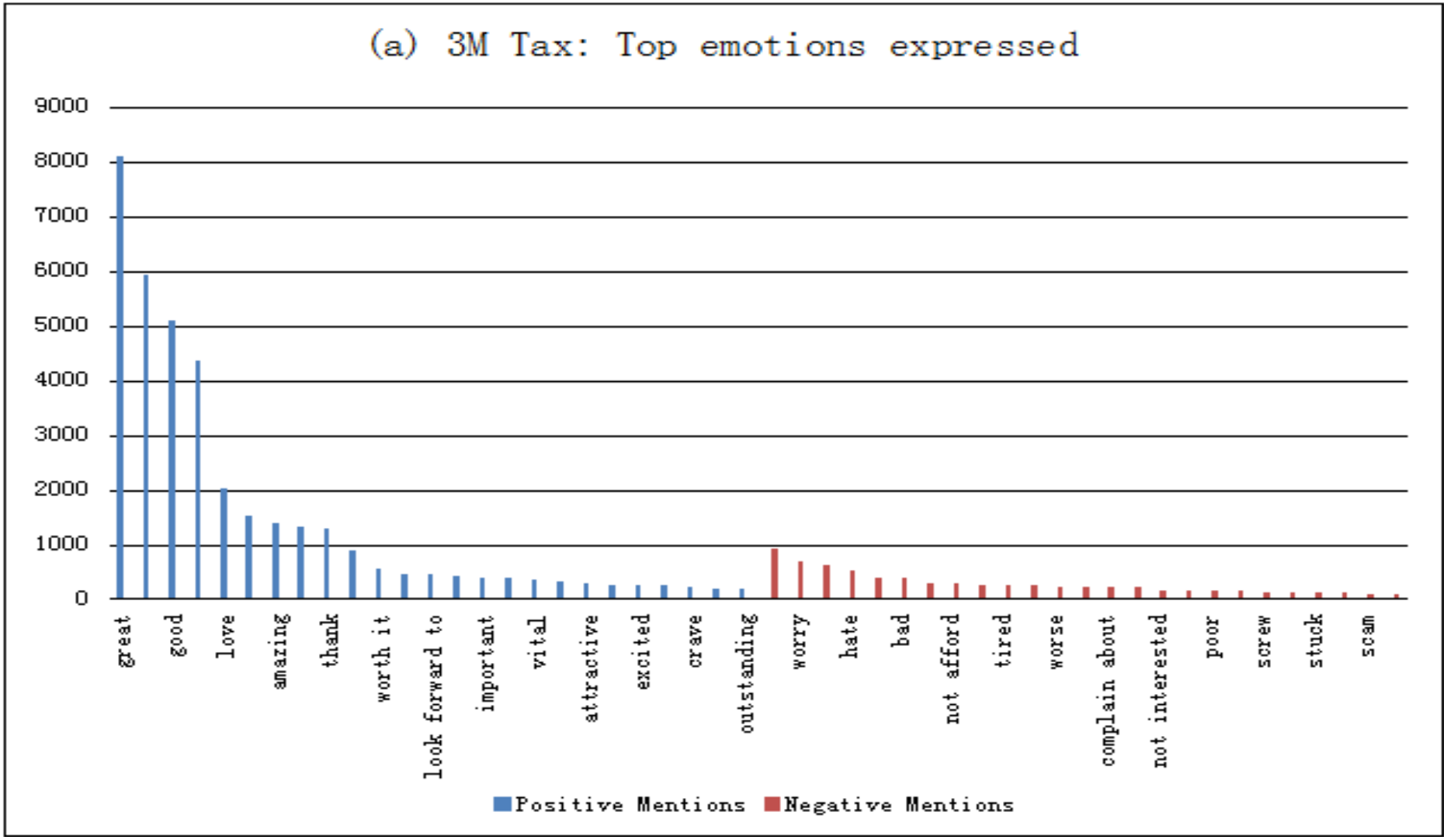
There is some concern regarding the efficacy of the data mining model of NetBase® and whether it is doing a good job of identifying the right contexts of conversations and extracting most of the relevant topics that relate to contextually linked keyword another concern is that data on social media networks is evasive of the data mining model assumptions used, given the pervasive use of diverse slang words in informal conversations.

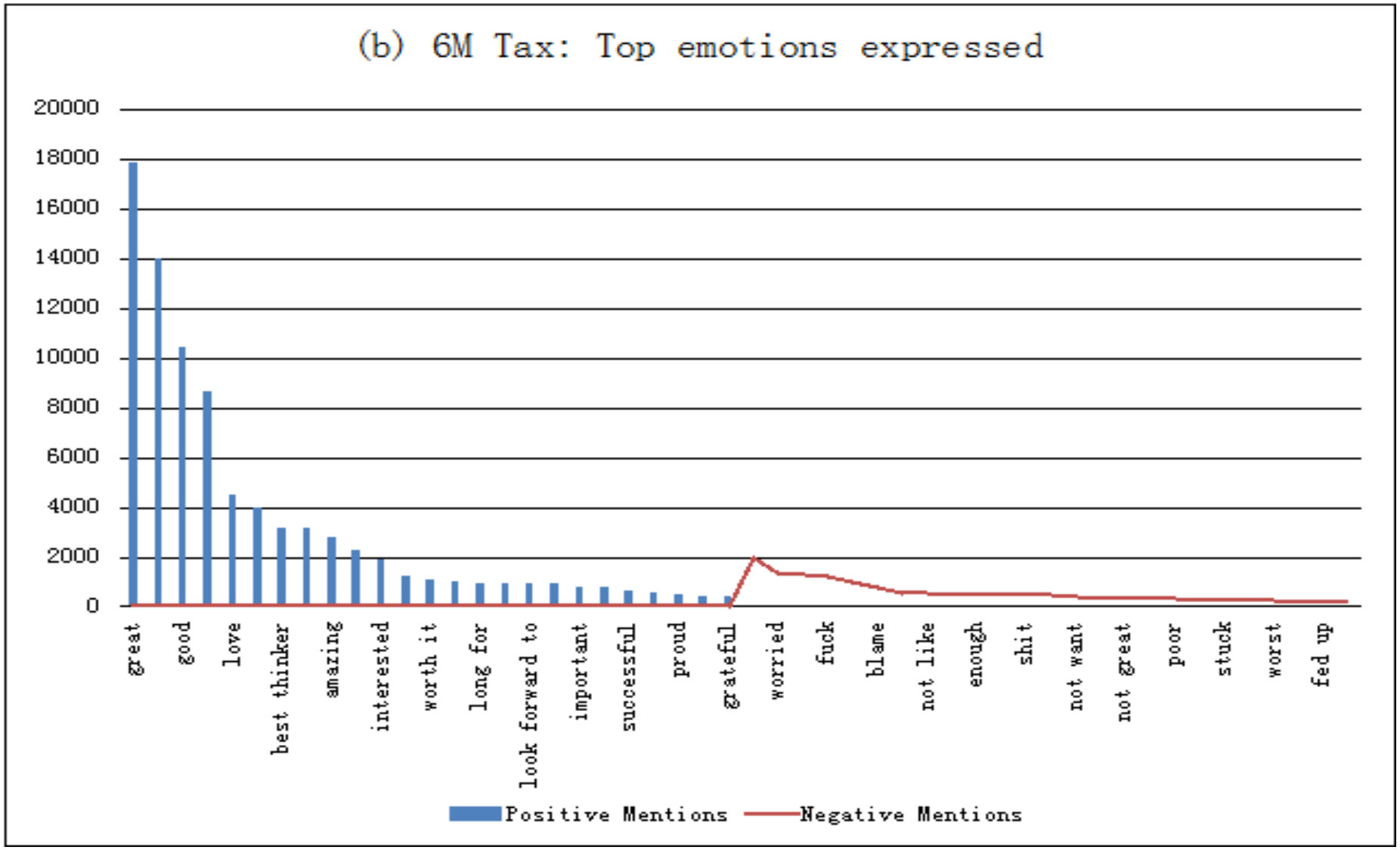
Furthermore, on the basis of the postulate that the main subject of discussion is about child benefits, just like in the qualitative analysis, the data does not tell whether the change made leads to a higher or a lower support of child benefits from the government. Ordinarily this would in turn give a hint of the likely direction of its impact on the economic situation of individuals. Therefore, in the absence of such link, it is appropriate to look at what is happening in the top emotions.

The top five emotions expressed: ‘great’, ‘good’, ‘love’, ‘amazing’ and ‘thank’ all point to a positive economic disposition about how people felt that year (Fig. 19). On the surface, the ‘top emotions’ seem to argue the point that whatever changes the government made to tax was a favorable one for families and consumers at large. Even on a basis of ratios of the most negative emotion expressed relative to the most positive emotion expressed for the whole year, this gives a ratio of 1:4 and implies that there is a negative emotion expressed for every four positive emotions expressed.

If the least five positive emotions were taken into account for the whole year (12-month period only), the words ‘favorite’, ‘proud’, ‘successful’, ‘important’ and ‘vital’ depict an organic picture of a year that has been highly favorable. Likewise the least five negative expressions for the whole year paint a different picture. However their frequency is lesser than those of their counterparts in the least five position emotions expressed. This implies overall, that people’s economic situation was better.

Tax: Top emotions expressed




(b) 6M Tax: Top emotions expressed
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
great
good
love
best thinker
amazing
interested
worth it
long for
look forward to
important
successful
proud
grateful
worried
fuck
blame
not like
enough
shit
not want
not great
poor
stuck
worst
fed up
Positive Mentions
Negative Mentions

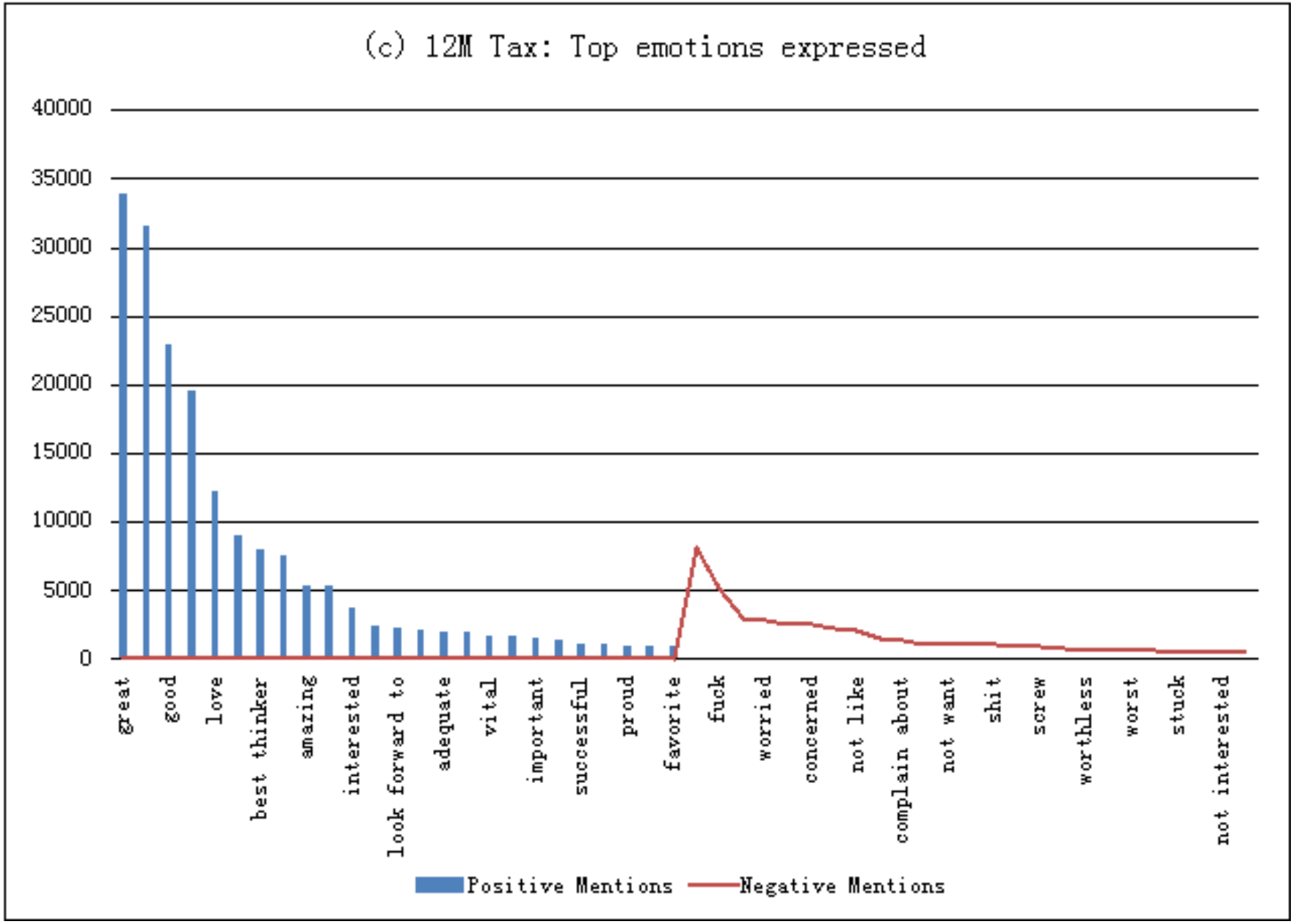


Fig. 19: Quantified representation of the top emotions expressed that are related to Tax

This observation is consistent with the global index of consumer confidence published by Nielsen for the year 2013. The Nielsen index results estimated there was a 1 percentage point increase in global consumer confidence (The Nielsen Company, 2014). This does give some credence to the opnion that social media networks data has a degree of correlation to other established measures of consumer confidence by renowned organizations like The Nielsen Company.

The other sub category, born in the era of social media networks—the hashtags gives another perspective to the story. Firstly, in comparison to normal vocabulary expressions, its use is still small. For example for the period 2012 – 2013, only 25000 of the most used hashtag, '#funding' occurred across social media networks. It must be noted that hashtags are used predominantly in Twitter, by its 200 million active monthly users (Jim Edwards - Business Insider, 2013).

In spite of the small scale representation of hashtags relative to the number of active users of Twitter, the five top hashtags of the year seem to reinforce the postulation that changes to government tax policies were economically beneficial. And a retrospective comparison between the least top five hashtags and the top five negative emotions supports drawing such conclusion.

So far the quantitative analysis of the NetBase® data of Table 2, graphed in Fig. 18, Fig. 19 and Fig. 20 illustrates that data from social media networks possess attributes that are good enough for deriving knowledge of higher order level. This is based on the observation that the data reflects directly what online conversations are about at each point in time.

Other findings the analysis has uncovered are that hashtags may not be the most effective tool for detecting what concerns people or is of most interest to them, when it comes to analysis of conversations on social media networks. The analysis also reflected shifts in online conversations through the movement of different top terms, as observed in Fig. 18, where it became obvious that tax policies affecting families changed.

Quantitatively, it is unclear why the volume of the number of terms does not scale with the user activity on the popular social media networks. But in the face of an assumption that the data mining model implemented in NetBase® is highly accurate, this reinforces claims that data on social media networks is noisy to an extent (Daas P.J.H and Puts M.J.H - European Central Bank, 2014).

At this point, further exploration of the attributes of the data from NetBase® is in order. In a similar line of thought as section 4.1.1, the data for Capital Expenditure will be analyzed in the following subsection.

Tax: Most used hashtags

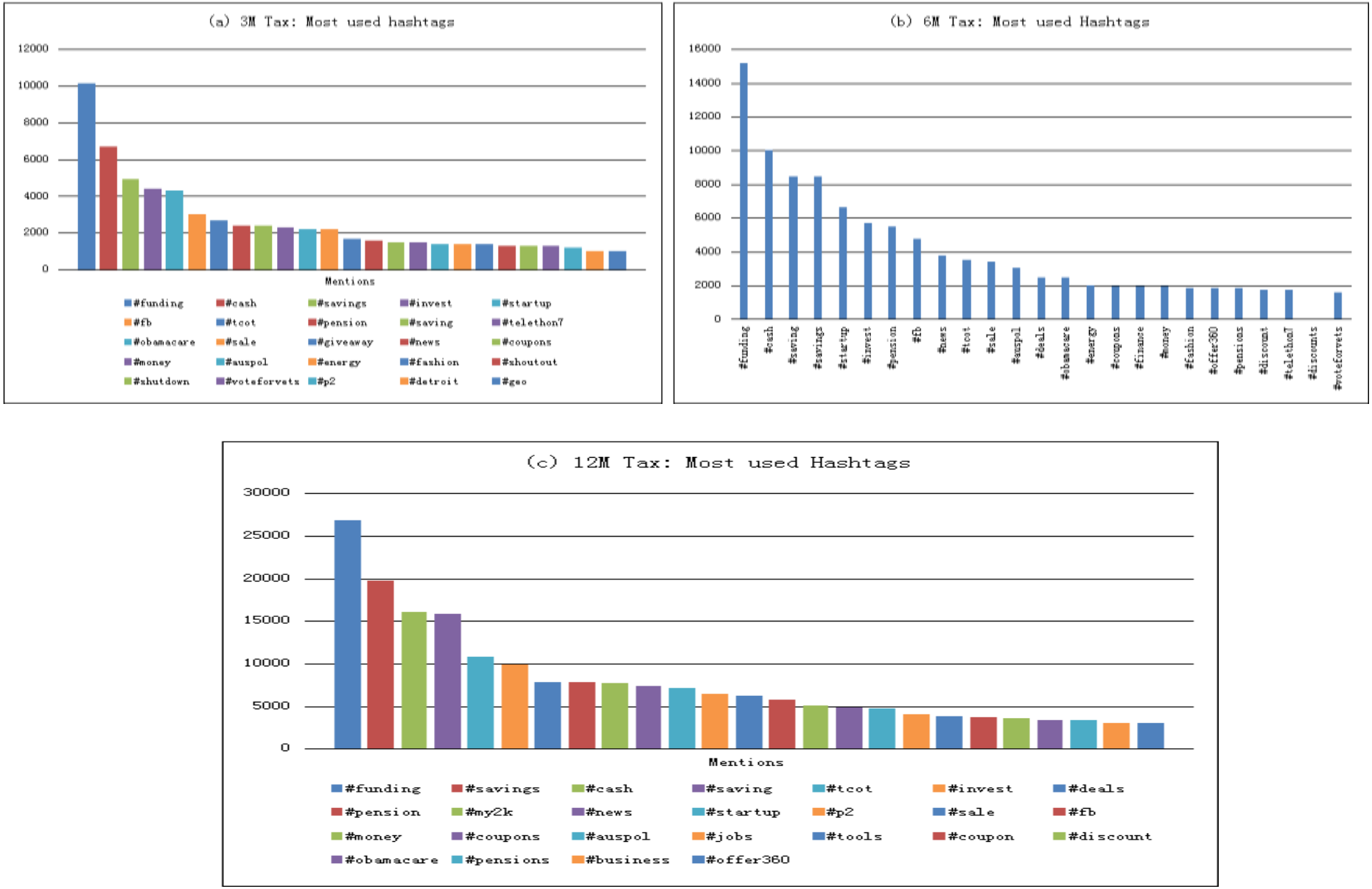


Fig. 20: The hashtags used in conversations that are contextually associated with Tax

Capital Expenditure

As an illustration, Fig. 21(c) (p. 72), shows that the five top terms of the year (12 month period), i.e. 'government', 'costs', 'market', 'tax' and 'economy' are same as the five top terms in the three month and six month periods, given they all relate to the economic operations of governance. But 'government' is by far the most discussed subject relative to the next top term of the year.

The numbers tell us that the capital expenditure keyword is not as conversational in informal contexts like the Tax keyword, owing to the frequency of the top term, 'government'. For instance 'government' was discussed 300,000 times less relative to the number of times it was discussed within the context of the Tax keyword for the whole of the year (12 month period). The five least top terms on the other hand do not repeat and so do not give a consistent picture if we look at the entire year's data.

In comparison to the volume of active users on the social media networks, based on Facebook's 1billion active monthly users, the number of times 'government' is discussed is a meager 0.08% of this number of users. Perhaps the fraction of users who actually post comments, messages, update their status would determine whether this number is indeed small.

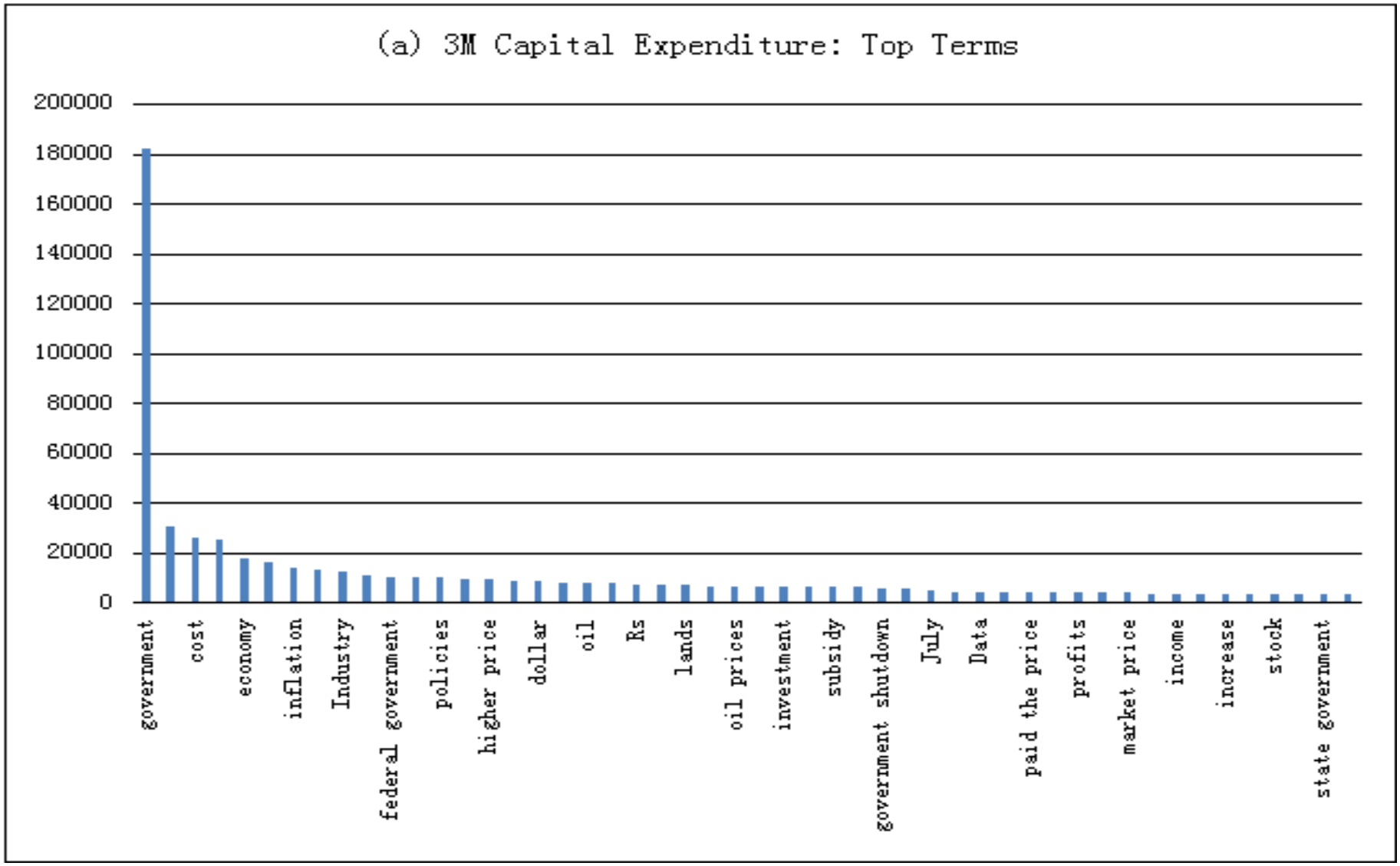
Otherwise, this is even more disappointing than that of tax, which happens to be 0.04% higher. And if an assumption that the data mining model error is insignificant, then this tells us that social media networks data is indeed noisy. This implies that there is less useful content than intuition makes us believe. Or that the number of users does not scale with the volume of active usage in form of messages, tweets or blogs. Since such data is not available, we would assume that posting of comments and online conversations would be up to 1% at least.

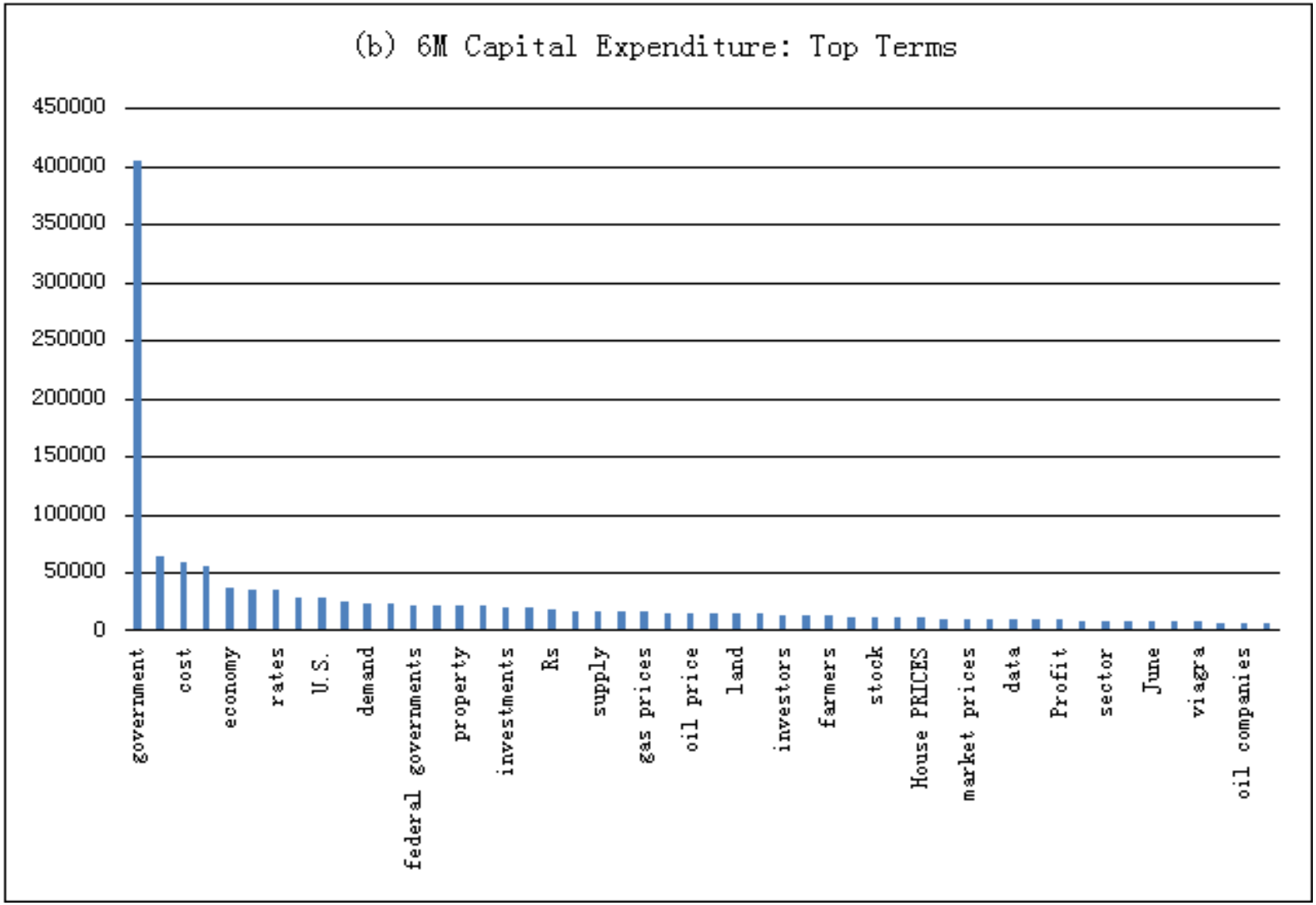
Looking into the 'top emotions expressed', Fig. 22(c) for the whole year, both volume and frequency is very small compared to the number of active users of Facebook's 1 billion active monthly users. As a ratio of the total number of active monthly users, this represents 0.00014%. Just like the qualitative analysis highlighted, there is no hint of how consumers feel. The sub category of the most used hashtags, Fig. 23 does not give this away also.

The analyzes show that the usefulness of data from social media networks rely on the data mining model of the data collection tool used to extract vocabulary associated with the context of interest. As a result, there is an assumption that the model embeds any causal relationships between the data relative to the knowledge domain of interest.

In the next section, the analysis is beamed on the standings of social media networks as sources of data. In particular the network from which much of the data originated.

Capital Expenditure: Top terms




(b) 6M Capital Expenditure: Top Terms
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
government
cost
economy
rates
U.S.
demand
federal governments
property
investments
Rs
supply
gas prices
oil price
land
investors
farmers
stock
House PRICES
market prices
data
Profit
sector
June
viagra
oil companies

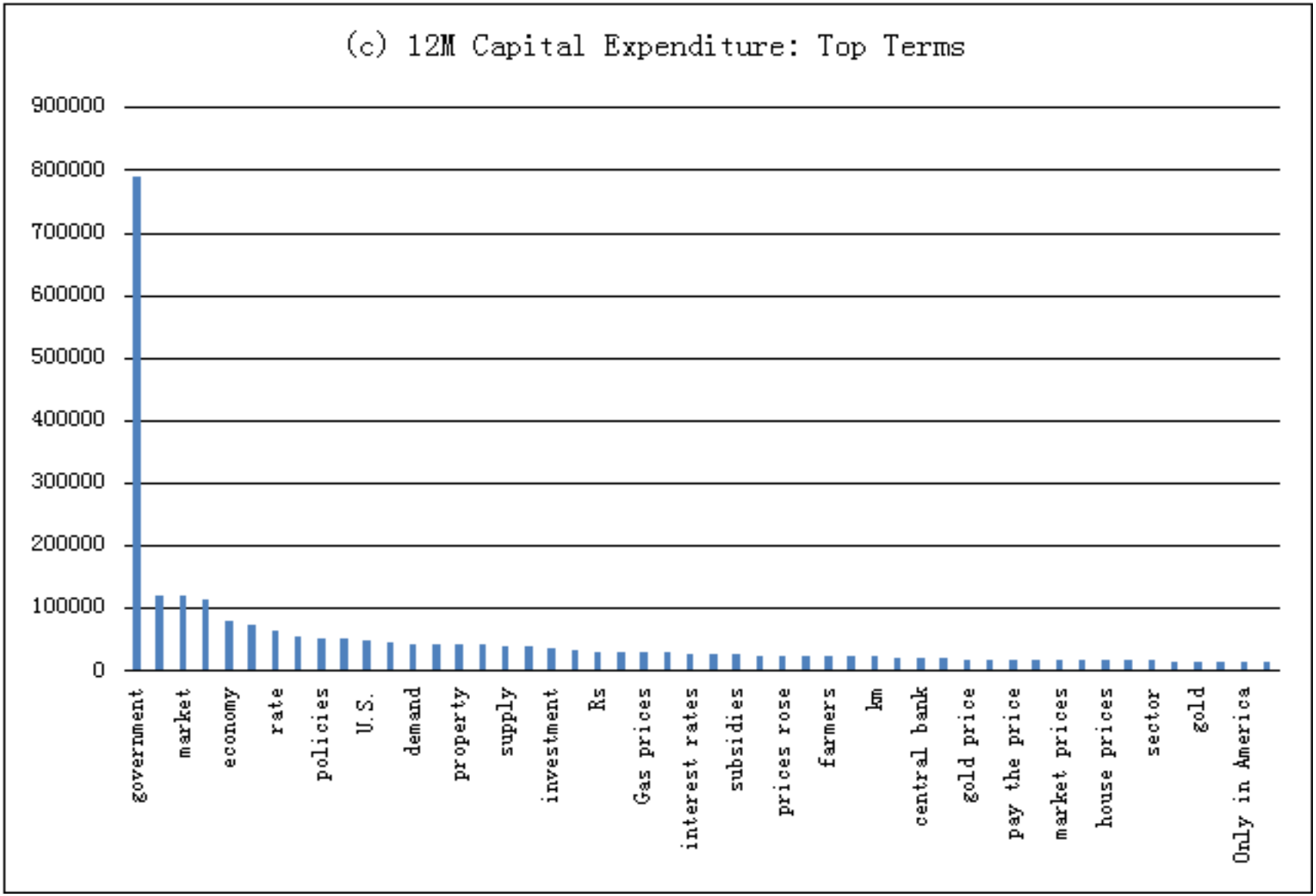
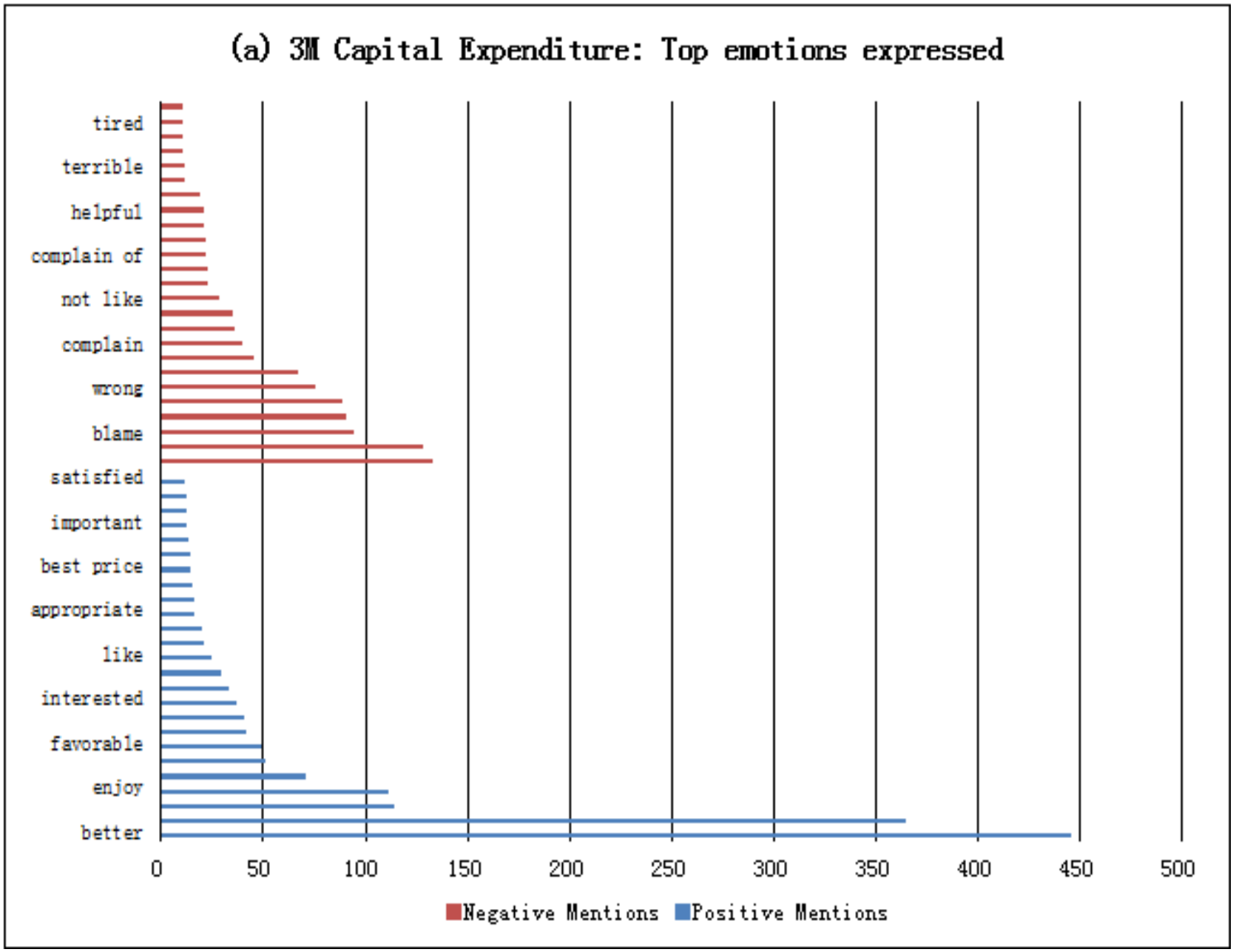
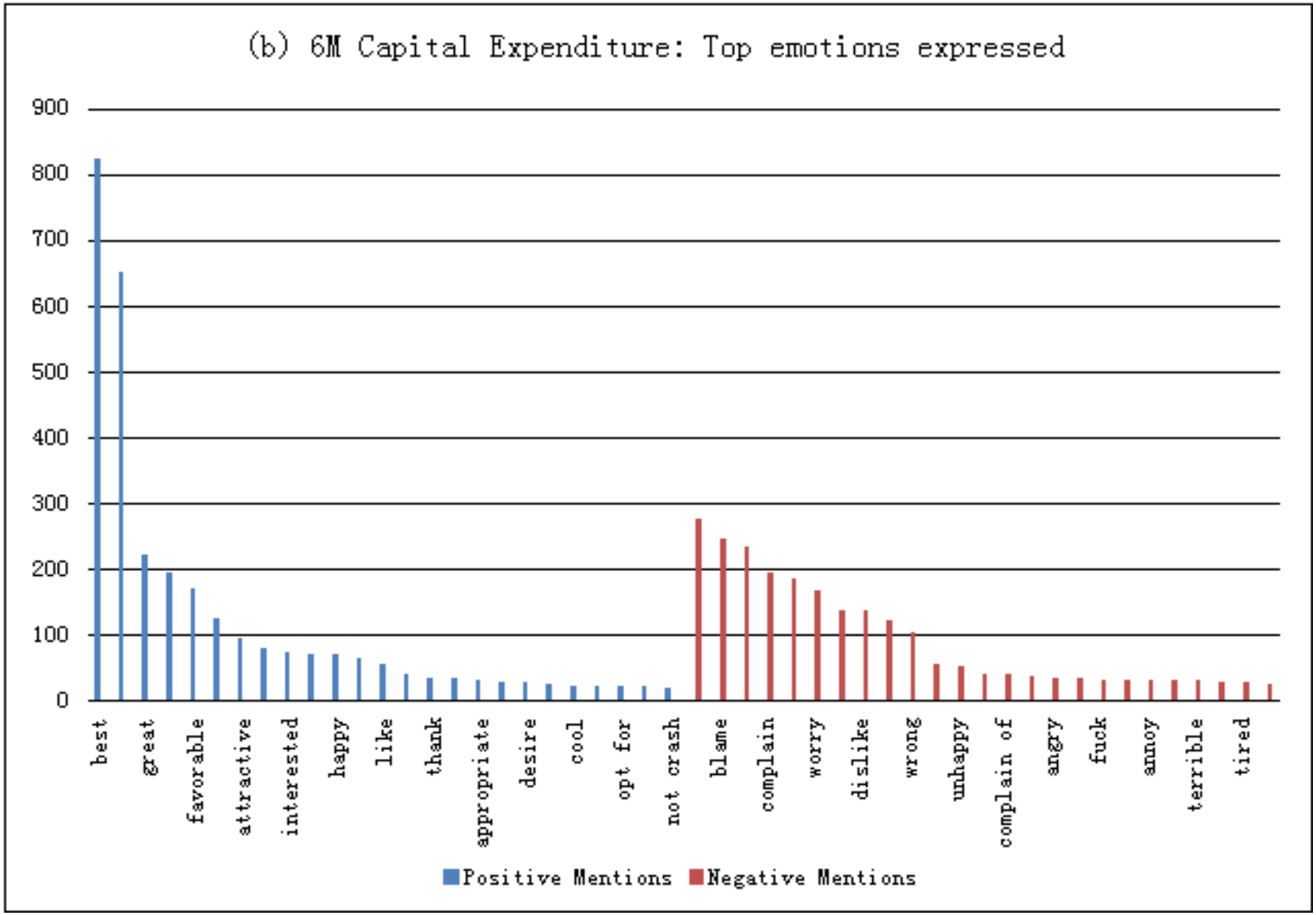


Fig. 21: The volume of top terms associated with Capital Expenditure

Capital Expenditure: Top emotions expressed




(b) 6M Capital Expenditure: Top emotions expressed
900
800
700
600
500
400
300
200
100
0
best
great
favorable
attractive
interested
happy
like
thank
appropriate
desire
cool
opt for
not crash
blame
complain
worry
dislike
wrong
unhappy
complain of
angry
fuck
annoy
terrible
tired
Positive Mentions
Negative Mentions

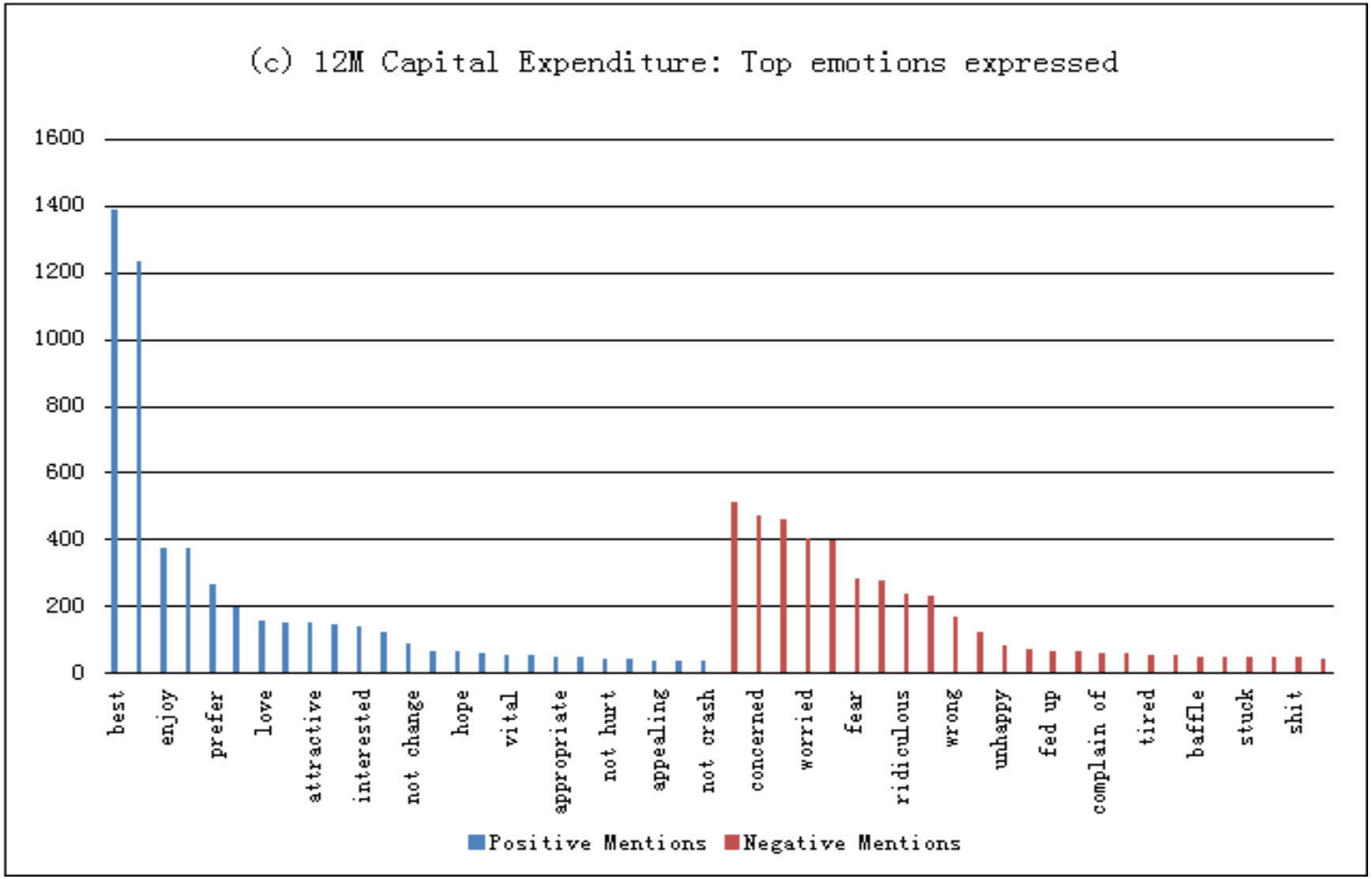


Fig. 22: Graphical representation of the number of Top emotions expressed, relative to Capital Expenditure

Capital Expenditure: Most used hashtags

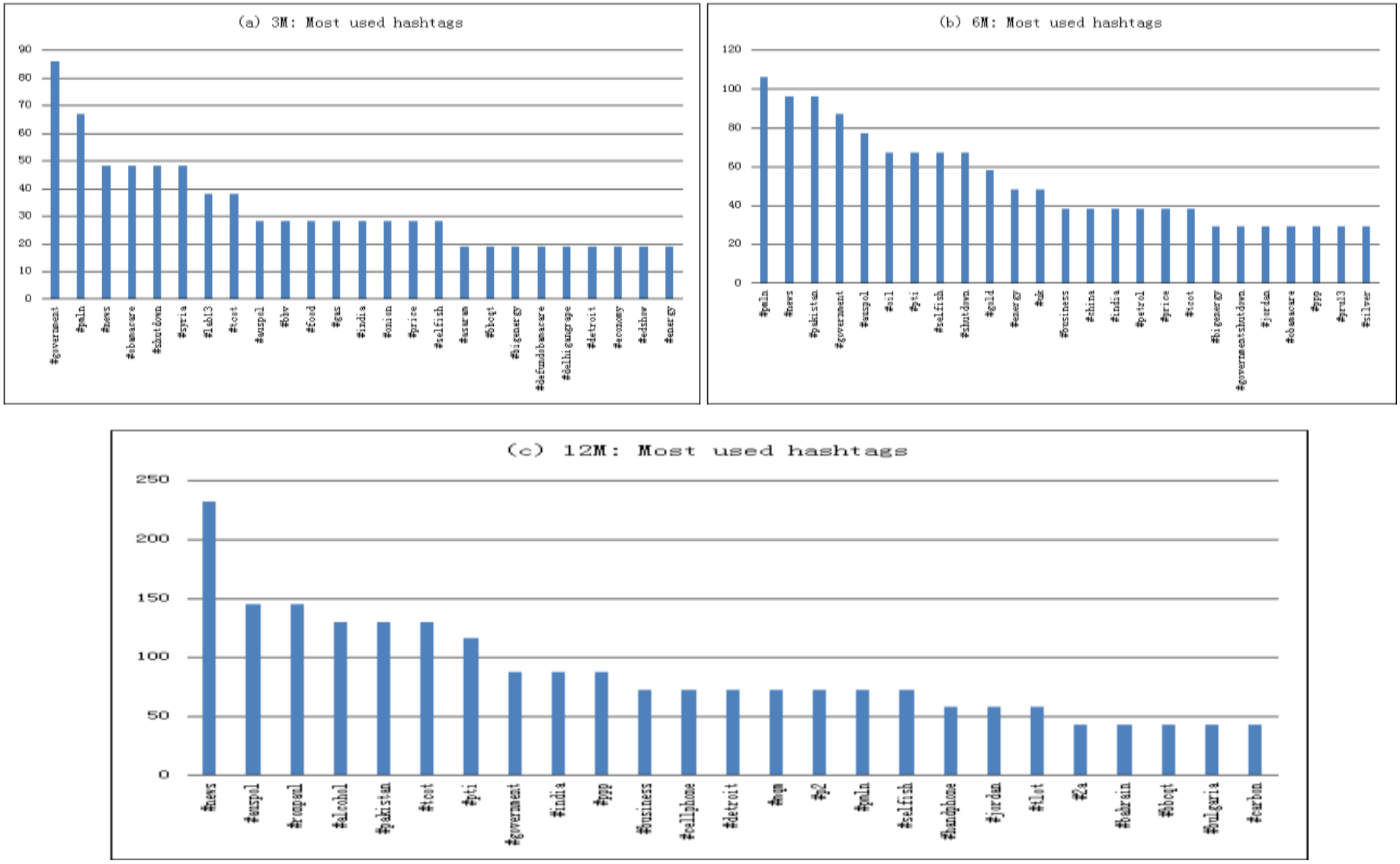


Fig. 23: The most used hashtags related to the subject of Capital Expenditure

4.1.3 Part C: An analysis of SMNs content contribution

The analysis done so far has ignored the social media networks, the sources of the data NetBase extracted for each keyword. The ensuing analysis looks at some these social media networks with the highest number of conversations that yielded the most relevant data. Then a quantitative analysis compares the number of conversations to the number of active users on the top two social media networks, to further exploit the attribute of 'noisiness' and data collection 'model error.'

Figures 25(A) – (G) for the 12 month period, show that Facebook is the social media network with the highest number of conversations because of the massive size of its network. The keyword Consumer Confidence is the only exception, where Twitter dominated. This outlier, in some way flags an attribute of social media networks data that suggests that the use of the data may best be served through a particular source or group of sources. Across the seven keywords Table 4 below shows the top five social media networks to buttress this observation.

Facebook.com, twitter.com, reddit.com and scribd.com are the top sources, as seen in Table 4. But their position in the five top conversation places show that some social media networks could be better sources with less noise. Of course this depends on the subject of interest. For instance, 'tax' presents every traveler with concerns to do with exchange rate or commodity prices. Businessmen and women whose businesses are international also face the reality of taxation wherever they do business.

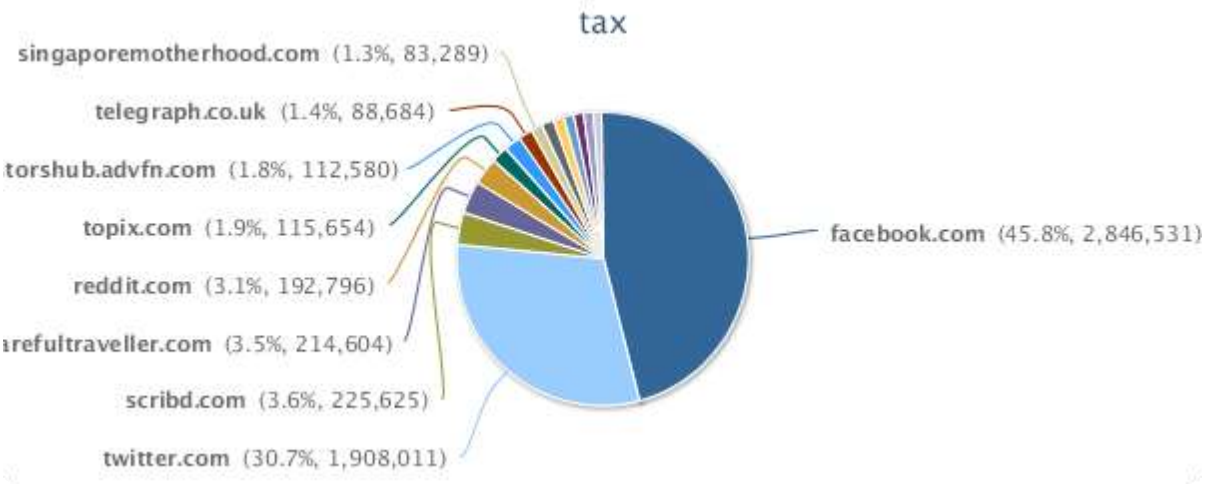
Table 4: Illustration of the how niche social media networks could serve some subject area better than the others

Keywords	**1st**	**2nd**	**3rd**	**4th**	**5th**
1. Tax	Facebook.com	Twitter.com	Scribd.com	Carefultraveller.com	Reddit.com
2. Capital Expenditure	Faceboook.com	Realestateforum.com	Twitter.com	Scribd.com	Reddit.com
3. Health	Facebook.com	Scribd.com	Twitter.com	Reddit.com	Huffingpost.com
4. Retirement	Facebook.com	Twitter.com	Topix.com	Reddit.com	Scribd.com
5. Consumer Confidence	Twitter.com	Facebook.com	Fontcraft.com	Investorshub.advnj.com	Online.wsj.com
6. Political Affiliation	Facebook.com	Twitter.com	Reddit.com	Topix.com	Scribd.com
7. Employment Condition	Facebook.com	Twitter.com	Scribd.com	Blogtalkradio.com	Reddit.com

realestateforum.com is a strong source of information in the subject of 'Capital Expenditure; Political Affiliation—topix.com; Employment condition—blogtalkradio.com and Retirement—topix.com. In fact the summary of Table 4 is that some social media networks may serve certain purposes better, given their privileged domain expertise on the subject of interest.

Note, however, that the data in table 4 and in figures 24 a-g does not take into account the number of users who could potentially be posting on each website, nor does it take into account the total number of messages posted. We did not have access to such data. If we had, it might have indicated for example that while carefultraveler.com has many fewer posts than Facebook about Tax, it might still be considered more influential based on the percentage of users posting or the percentage of posts related to Tax.

Thinking of the fourteen keywords used to extract data from social media networks, it turns out that a social media network such as Twitter is favorable to the use of phonemes given the 140 character restriction on the length of a tweet (Twitter Inc., 2014). Whereas an attempt to determine the character limit on Facebook shows that upwards of 1000 characters is permitted. This implies that a data mining model that is based on single syllable words would easily process content from Twitter given its briefness than say scribd.com, which is a social network for electronic books by unpublished authors. This is because the longer the content, the higher the chances of inaccurately deducing the intentions and actual context of the conversation.

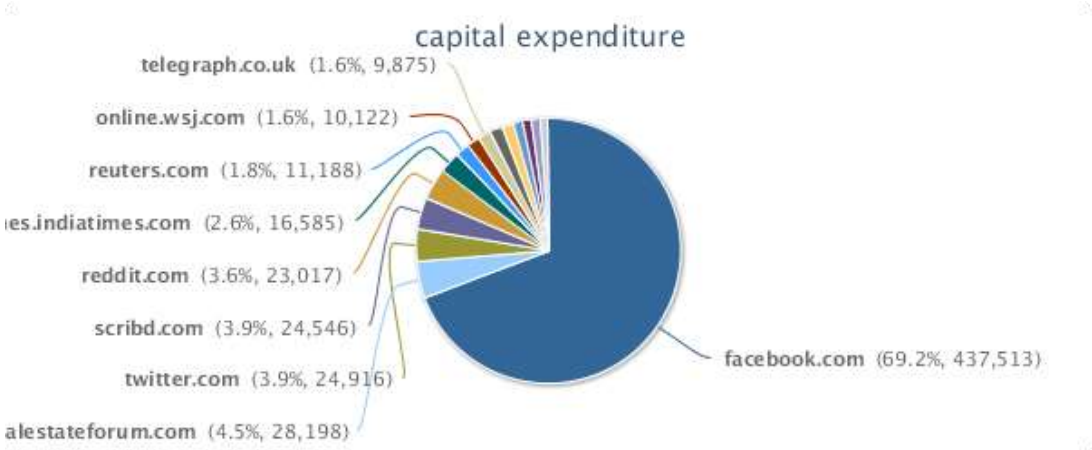


(a)

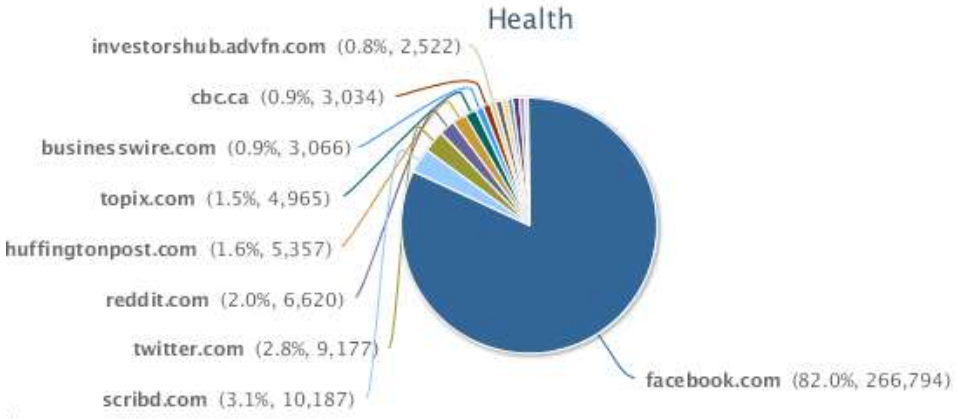
Nonetheless, the data shows that Facebook is a dominant influencing source in social media. This is in part because Facebook permits the posting of story long posts (the length of blogs), short Twitter like posts and comments by any of its over one billion users. These seem to work in its favor compared to the other social media networks.

Another observation is that the keywords themselves and where most of their data has come from matters. It is tempting to say that single syllable keywords are best sourced from a social media network like Twitter and secondary words from other websites that permit a lot of text. This theory could be supported by looking at Fig. 24(a) for Tax, but is debunked by observing that Facebook is still the major source for other single syllable keywords like (c) Health and (d) Retirement.

In addition, one could say that keywords that are used frequently in formal conversations about the economy tend to be announced in the news. As such they are most likely going to get attention on Facebook and Twitter, the major networks that also supports 'announcement' conversations. On the other hand a keyword that has a perspective of 'self-interest/learning' to it is most likely to be found on networks where there are lengthy, detailed and informed conversations. This is particularly the case with the keyword 'Health'. As seen in Fig. 24(c), scribd.com, the online self-publishing service of e-books is the second highest source after facebook.com (notwithstanding the difference in base audience numbers).



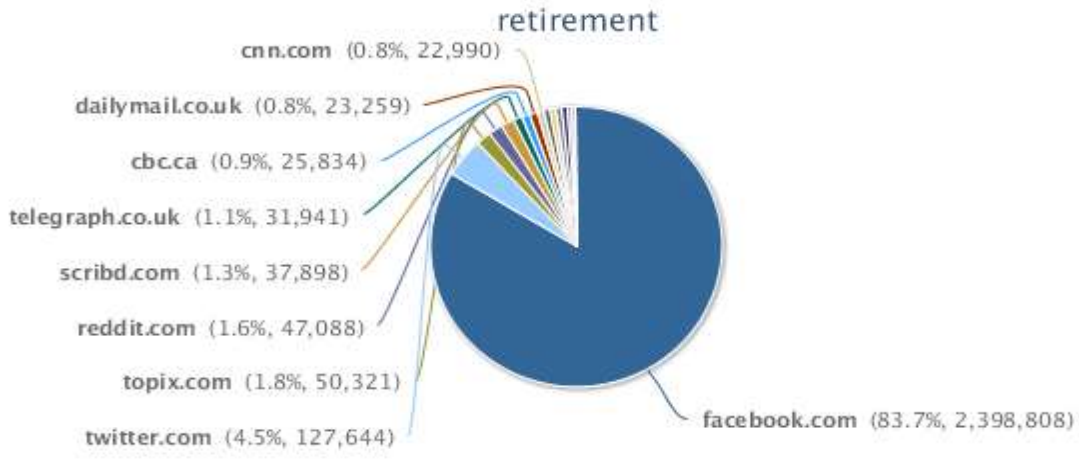
(b)



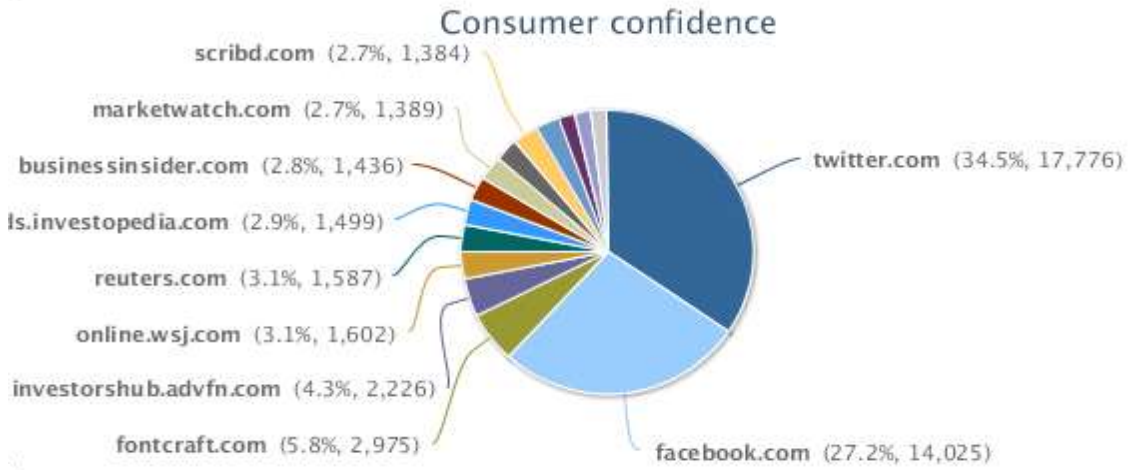
(c)

An analysis of the numbers shows other hidden perspectives about social media network users and what they feel happy to share. Note that Fig. 24(a) – (g) gives a quantified measure of the number of conversations that contain each keyword, as opposed to the number of online conversations that took place on each social media network. On this basis, it could be claimed that topics that are personal and private are less talked about that those that are external to an individual.

Health (c) and Consumer Confidence (e), seem to be a private affair when compared to Retirement (d) and Political Affiliation (f). If the top conversation source is considered, Health (c), for example was found in 270,000 conversations on Facebook compared to the 2.3 million on Political Affiliation or 2.8 million on Tax. More so, Facebook being a very informal network where there are closer interactions between family and friends, held the most conversations on Health of 270,000, in comparison to the second place network, scribd.com containing 10,000, even though scribd.com is a prolific [electronic] book open publishing service with 60 million documents (Scribd., 2015).

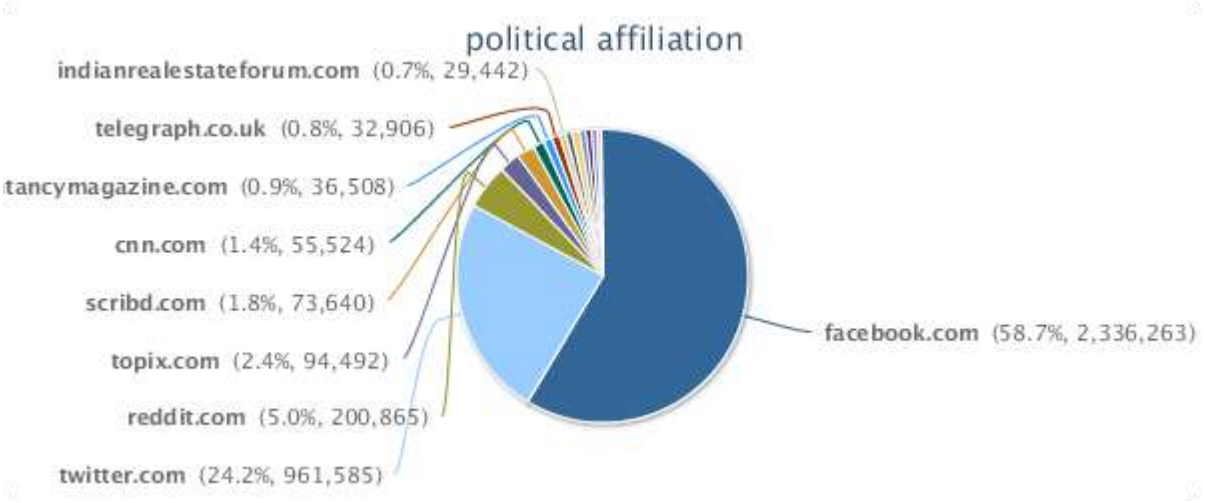


(d)

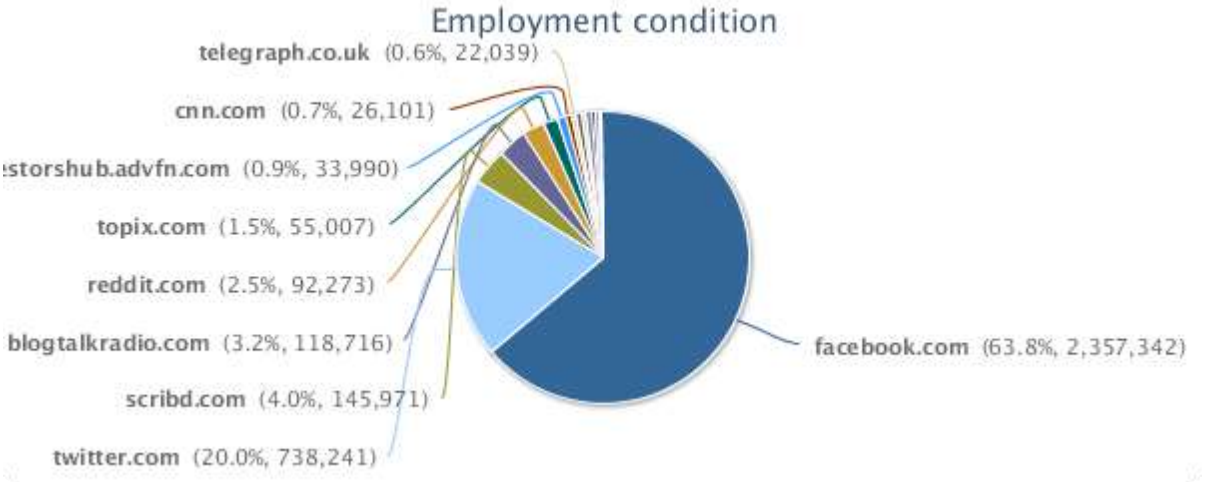


(e)

The numbers associated with the number of conversations held on SMNs shed some light on the degree of how noisy data from SMNs is. In 2014 it was claimed that 510 comments and 293,000 statuses were posted on Facebook every second (Dan Noyes - Zephora Internet Marketing Solutions, 2014). This results to a total of ~300,000 conversations per second. Multiplying this number by 60 gives 18 million conversations per hour. When this is multiplied by 24, it gives an estimation that about 432 million conversations per day. If night time usage and other cyclic causes of inactivity are taken into account, this number could be say 400 million per day. This means that for a 365-day year, approximately 150 billion conversations were posted on Facebook.



(f)



(g)

Fig. 24: The top social media networks in terms of the volume of conversations

Based on the results of Fig. 24 (a) – (g) which is for the period of 1 year—Oct. 2012 – Oct. 2013, there is a maximum of 3 million mentions of any one of our keywords on Facebook. This represents 0.002% of the total number of conversations had on Facebook in one year.

This implies one or more realities as follows: that the conversations on social media networks appear to be about other subjects that do not in any way relate to the economic situation of users. This hint of "irrelevance" is questionable because social media networks are used also by governments and businesses to solicit [subjective] commentary from people. However, caution must be exercised because this does not imply that the commentary responses are relevant to the topic of CC. Moreover it could be a consequence of the NLP algorithms not being sufficiently equipped with language rules that cater for the unstructured language usage on social media network conversations.

Moreover, there is the possibility that the rules and assumptions of the data mining model may not be aligned with the innate structure of data on social media networks. So in as much as the data may be noisy, it may also be down to data mining model error. Therefore, it calls to question the need to advance models that are tenable for deriving higher order levels of knowledge, beyond the literal extraction of phonemes based on foundational language rules.

At this point in the analysis, the data from social media networks have been explored qualitatively and examined quantitatively. The sources of the data have also been analyzed and shed insight into some opportunities for improvement in the methodologies of data mining models. However, there is a case for examining the characteristics of the raw data that NetBase mined as shown in Table 3(The third output from a NetBase search; the indexed dataset extracted from social media networks for the search term of Capital Expenditure.)

4.1.4 Part D: Analysis of raw data indexed by NetBase

The modus operandi of NetBase is that it indexes an enormous amount of data from across social media networks, thereby making these data available for research at any time. This dictates a reliance on the model used to mine social media networks for data, which is based on the concept of Natural Language Processing (NetBase, 2014). Also other factors being relied upon are the frequency of data collection, the metadata about the data that is indexed and the volume of the data it has to process.

All of the above have a fundamental effect on the data that have been analyzed so far. For instance it affects the accuracy of the classification of the words associated with the keywords concerning consumer confidence. The measured frequency of their occurrence, whether it is relevant in context and in meaning or not, and finally the likelihood of an amplified occurrence due to a high number of repeated raw data

being captured more than once due to a short interval between scans of social media networks for data all influences the usefulness and significance of the data.

Most of the attributes mentioned above can be seen in Table 3 on page 43. The actual posts, tweets or comments on social media networks are indexed under the column called Sound Bite Text and their classification as being either negative or positive as Sound Bite Sentiment. If we considered the negative attribute of repetition it can be seen that rows 10 – 11, 15 – 21 and 13 – 14 of Table 3 are the same. This represents a total duplication of 7 out of 30 indexed datasets, ~23% noise. For the sake of exploration, let's assume that this noise scaled up to the 500 raw datasets, i.e. that 23% of the dataset are repetitions. If we take the claim of the number of conversations on Facebook per hour as our benchmark, to be 300,000, it implies 70,000 of those 'conversations' are just repetitions not actual conversations on social media networks. This does throw concern and points to an opportunity for improvement of data mining models.

Furthermore, on the subject of contextual relevance, Table 5 gives a preview into the challenges of automated classification based on Natural Language Processing. Between the rows of 464 – 485, the raw sound bite is nothing short of a marketing article espousing the great deal vouchers on Amazon. In no way do these represent the sentiment of consumers who have used the vouchers and saved some money in the process. Instead they are just repeated marketing descriptions of vouchers offering good deals. This is the type of "noise" that was mentioned earlier, and it could be removed to filter the results and improve the accuracy of the application using the data.

The last attribute of frequency of sampling is examined using Table 6. It is obvious that the sampling frequency is anything from milliseconds to seconds and sometimes the lower tens of minutes. This is an incredibly short time owing to the slow pace of time at which human trends evolve, even though human events happen at a much faster pace. It could be said that this presents opportunities for repetitions but also for a noisy data that will require filters to get the real data. Nonetheless there is a case for oversampling in order to have a large enough sample that increases the presence of actual trends.

The case of oversampling is supported by proven concepts from the knowledge domain of Signal processing in Engineering, under a theory known as The Nyquist-Shannon's sampling theorem (Weisstein, Eric W. - From MathWorld -- A Wolfram Web Resource, 2015). It claims that in order to reconstruct a signal, in this case human expression of sentiment, that all available data on that signal must be sampled at twice the frequency of the signal of interest. In order words, to measure human events that occur within seconds, we must sample social media networks at least twice every second to get a picture of the actual events. Implicitly this would mean that noise will be present in the data as we have already explored.

Table 5: Indexed raw data from NetBase associated with Tax for 12 month period

Sequence	Sound Bite Text	Sound Bit
460	Disposable Blue Underpad Chux 23 X 36 150 case deals, coupons and promotions. Below you will get the best cost savings from Amazon.Take a look at links bellow to order this great product. Dis	Neutral
461	Cameron's Chocolate Caramel Brownie Single Serve Coffees, 12-Count deals, coupons and promotions. In this article you can find the perfect cost savings from Amazon.Have a look at links bellow	Neutral
462	Nippon Kodo - Kayuragi - Sandalwood 40 Sticks deals, coupons and promotions. The following you can find the perfect cost savings from Amazon.Check out links bellow to buy this wonderful prod	Neutral
463	iLIVING Infrared Portable Space Heater with Dual Heating System, 1500W, Remote Control, Dark Walnut Wooden Cabinet deals, coupons and promotions. Here you will get the perfect cost savings	Neutral
464	Francois et Mimi 14-Ounce Colored Ceramic Coffee/Soup Mugs, Large, Solid, Set of 6 deals, coupons and promotions. Right here you can get the perfect savings from Amazon.Take a look at links be	Positive
465	Zenith Premium Bathtub and Shower Pole Caddy, White deals, coupons and promotions. The following you can get the best cost savings from Amazon.Have a look at links bellow to pay for this grea	Neutral
466	Resco Deluxe Dog Nail Trimmer with Handle Grips, Large, Yellow deals, coupons and promotions. The following you will get the best savings from Amazon.Check out links bellow to order this good	Neutral
467	Solar Panel Cable 50 Ft - Mc4 Pv Extension- 10awg - 600vdc - Sunlight Resistant deals, coupons and promotions. Here you will find the greatest cost savings from Amazon.Take a look at links bellov	Neutral
468	Sprite SLC Replacement Slim Line Shower Filter Cartridge deals, coupons and promotions. The following you can get the perfect savings from Amazon.Have a look at links bellow to purchase this aw	Positive
469	Pinzon Ultrasoft 300 Thread Count Percale Sheet Set, Twin XL, Platinum deals, coupons and promotions. In this article you will get the perfect savings from Amazon.Have a look at links bellow to pu	Positive
470	CARL G-01 Replacement Straight Blade deals, coupons and promotions. The following you will find the greatest cost savings from Amazon.Check out links bellow to buy this good product. CARL G-0	Neutral
471	Kikkerland Rubber Ducky Tub Treads, Set of 5 deals, coupons and promotions. Right here you can find the best savings from Amazon.Check out links bellow to buy this wonderful product. Kikkerlar	Neutral
472	Beautyrest Firm Standard Size Twin Pack Bed Pillow deals, coupons and promotions. In this article you can find the perfect savings from Amazon.Check out links bellow to purchase this great prod	Positive
473	Breville BCI600XL Smart Scoop Ice Cream Maker deals, coupons and promotions. In this article you can find the best savings from Amazon.Check out links bellow to pay for this excellent product.	Neutral
474	Hydro Tools 8110 Weighted Half Moon Pool Vacuum Head deals, coupons and promotions. Right here you will find the greatest cost savings from Amazon.Have a look at links bellow to purchase t	Neutral
475	Clover Felting Needle Refill Fine Weight deals, coupons and promotions. Here you can find the best savings from Amazon.Take a look at links bellow to buy this excellent product. Clover Felting Ne	Neutral
476	High Cotton Moustache International Doormat deals, coupons and promotions. Below you will get the best savings from Amazon.Have a look at links bellow to buy this great product. High Cotton	Neutral
477	Dogs Playing Poker Metal Tin Sign deals, coupons and promotions. Right here you will get the perfect savings from Amazon.Have a look at links bellow to purchase this wonderful product. Dogs Pl	Positive
478	Sephra Premium Milk Chocolate deals, coupons and promotions. In this article you can find the perfect savings from Amazon.Have a look at links bellow to buy this good product. Sephra Premium	Positive
479	Hayward ECX1040 Replacement Pool Filter Bump Handle (EC40) deals, coupons and promotions. The following you can find the greatest cost savings from Amazon.Have a look at links bellow to de	Neutral
480	Breville BOV800PS13 13-Inch Pizza Stone for use with the BOV800XL Smart Oven deals, coupons and promotions. Below you can get the best savings from Amazon.Take a look at links bellow to buy	Neutral
481	HAAN RMF-2X Ultra-Clean Pads, Ultra-Microfiber Steam Cleaning Pads For All HAAN FS, SI and MS series steamers; 2 Pack deals, coupons and promotions. In this article you will find the best savir	Neutral
482	Mc Gill Paper Blossom Tool Kit 4/Pkg deals, coupons and promotions. Here you will find the perfect savings from Amazon.Check out links bellow to buy this good product. Mc Gill Paper Blossom T	Positive
483	Eight O'Clock Coffee, Dark Italian Roast Whole Bean, 11.5-Ounce Bag (Pack of 4) deals, coupons and promotions. Below you can get the best savings from Amazon.Have a look at links bellow to dec	Neutral
484	Hopkins 47235 Impulse Brake Control deals, coupons and promotions. In this article you will get the perfect savings from Amazon.Have a look at links bellow to buy this amazing product. Hopkins	Positive
485	Night Festival (Blue) deals, coupons and promotions. Below you can get the perfect cost savings from Amazon.Take a look at links bellow to buy this great product. Night Festival (Blue) BIG IMAGE	Neutral
486	Tommy Hilfiger Preppy Palm Beach Towel deals, coupons and promotions. In this article you can get the best savings from Amazon.Have a look at links bellow to decide to buy this excellent produ	Neutral
487	Fresh Flowers - Purple Dendrobium Orchids with Vase deals, coupons and promotions. Right here you can find the perfect savings from Amazon.Check out links bellow to pay for this great product	Positive
488	Strathwood Griffen All-Weather Wicker 3-Seater Sofa, Dark Brown deals, coupons and promotions. Right here you will get the perfect savings from Amazon.Take a look at links bellow to purchase t	Positive
489	Vera Wang by Wedgwood With Love Cake Knife and Server deals, coupons and promotions. Right here you will get the best savings from Amazon.Have a look at links bellow to pay for this good pro	Neutral
490	Wooden Wine Rack-Holds 44 Bottles-Unfinished Pine (Unfinished Pine) (40.5"h x 17"w x 10.5"d) deals, coupons and promotions. Right here you can get the perfect cost savings from Amazon.Have	Neutral
491	Andrew Wyeth Benner Island, Maine deals, coupons and promotions. Right here you will get the perfect savings from Amazon.Have a look at links bellow to decide to buy this amazing product. An	Positive
492	Taylor 9867FDA Digital Thermocouple Thermometer with Folding Probe deals, coupons and promotions. The following you will find the perfect savings from Amazon.Take a look at links bellow to c	Positive
493	Smelleze® Reusable Formaldehyde Absorbent Pouch: XX Large - Treats 300 Sq. Ft. deals, coupons and promotions. The following you will find the best savings from Amazon.Check out links bellow to	Neutral
494	All with a 100% Lifetime Guarantee to make sure you are Completely Satisfied!! (4) deals, coupons and promotions. The following you can find the greatest savings from Amazon.Have a look at link	Neutral
495	Dogit Fountain Pump Replacement for Dog Drinking Fountain deals, coupons and promotions. Right here you can get the perfect cost savings from Amazon.Take a look at links bellow to order this e	Neutral
496	Grenade Ice Cube Tray deals, coupons and promotions. In this article you will find the best cost savings from Amazon.Take a look at links bellow to purchase this awesome product. Grenade Ice C	Neutral
497	doTerra Elevation Essential Oil Blend 15 ml deals, coupons and promotions. Right here you will get the best savings from Amazon.Check out links bellow to buy this excellent product. doTerra Elev	Neutral
498	Winsome Wood 5-Piece TV Table Set, Natural deals, coupons and promotions. In this article you will find the perfect cost savings from Amazon.Have a look at links bellow to pay for this excellent	Neutral
499	Wellness 3-Ounce Minced Salmon Canned Cat Food, Pack of 24 deals, coupons and promotions. The following you can get the best savings from Amazon.Have a look at links bellow to order this go	Neutral
500	Winix WAC9500 Ultimate Pet True HEPA Air Cleaner with PlasmaWave Technology deals, coupons and promotions. The following you can get the best cost savings from Amazon.Have a look at links	Neutral

Table 6: Part B of Table 5 showing the sampling interval

456	Positive	Offer – Fen	Blogs	http://infocheap.info/offer-fenix-d	infocheap.info	2013-08-07 20:34			Anonymou				Yes
457	Neutral	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 20:33:18			Anonymou				Yes
458	Neutral	Buy – Pet-T	Blogs	http://infocheap.info/buy-pet-tinic	infocheap.info	2013-08-07 20:33:12			Anonymou				Yes
459	Neutral	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 20:24:42			Anonymou				Yes
460	Neutral	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 20:24:03			Anonymou				Yes
461	Neutral	Cheap + Ca	Blogs	http://infocheap.info/cheap-camer	infocheap.info	2013-08-07 20:23:57			Anonymou				Yes
462	Neutral	Offer ! Nip	Blogs	http://infocheap.info/offer-nippon	infocheap.info	2013-08-07 20:23:03			Anonymou				Yes
463	Neutral	Offer $ iLI\	Blogs	http://infocheap.info/offer-iliving-	infocheap.info	2013-08-07 20:14:24			Anonymou				Yes
464	Positive	Buy ? Fran	Blogs	http://infocheap.info/buy-francois	infocheap.info	2013-08-07 20:13:42			Anonymou				Yes
465	Neutral	Offer & Zer	Blogs	http://infocheap.info/offer-zenith-p	infocheap.info	2013-08-07 20:08:15			Anonymou				Yes
466	Neutral	Buy – Resc	Blogs	http://infocheap.info/buy-resco-de	infocheap.info	2013-08-07 19:56:18			Anonymou				Yes
467	Neutral	Cheap + So	Blogs	http://infocheap.info/cheap-solar-	infocheap.info	2013-08-07 19:54:36			Anonymou				Yes
468	Positive	Cheap ? Sp	Blogs	http://infocheap.info/cheap-sprite	infocheap.info	2013-08-07 19:54:09			Anonymou				Yes
469	Positive	Buy # Pinz	Blogs	http://infocheap.info/buy-pinzon-u	infocheap.info	2013-08-07 19:49:33			Anonymou				Yes
470	Neutral	Offer # CAF	Blogs	http://infocheap.info/offer-carl-g-(	infocheap.info	2013-08-07 19:48:30			Anonymou				Yes
471	Neutral	Buy . Kikke	Blogs	http://infocheap.info/buy-kikkerla	infocheap.info	2013-08-07 19:47:30			Anonymou				Yes
472	Positive	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 19:47:27			Anonymou				Yes
473	Neutral	Buy $ Brev	Blogs	http://infocheap.info/buy-breville-	infocheap.info	2013-08-07 19:42:45			Anonymou				Yes
474	Neutral	Offer – Hyc	Blogs	http://infocheap.info/offer-hydro-t	infocheap.info	2013-08-07 19:35:54			Anonymou				Yes
475	Neutral	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 19:35:48			Anonymou				Yes
476	Neutral	Cheap : Hig	Blogs	http://infocheap.info/cheap-high-c	infocheap.info	2013-08-07 19:35:27			Anonymou				Yes
477	Positive	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 19:33:24			Anonymou				Yes
478	Positive	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 19:29:09			Anonymou				Yes
479	Neutral	Offer % Ha	Blogs	http://infocheap.info/offer-haywar	infocheap.info	2013-08-07 19:24:21			Anonymou				Yes
480	Neutral	Buy ' Brevi	Blogs	http://infocheap.info/buy-breville-	infocheap.info	2013-08-07 19:22:36			Anonymou				Yes
481	Neutral	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 19:15:24			Anonymou				Yes
482	Positive	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 19:15:09			Anonymou				Yes
483	Neutral	Cheap ? Eig	Blogs	http://infocheap.info/cheap-eight-	infocheap.info	2013-08-07 19:14:39			Anonymou				Yes
484	Positive	Offer : Hop	Blogs	http://infocheap.info/offer-hopkin	infocheap.info	2013-08-07 19:14:15			Anonymou				Yes
485	Neutral	Buy ' Night	Blogs	http://infocheap.info/buy-night-fes	infocheap.info	2013-08-07 19:01:57			Anonymou				Yes
486	Neutral	Cheap ' Toi	Blogs	http://infocheap.info/cheap-tomm	infocheap.info	2013-08-07 18:59:48			Anonymou				Yes
487	Positive	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 18:54:57			Anonymou				Yes
488	Positive	Cheap @ S	Blogs	http://infocheap.info/cheap-strath	infocheap.info	2013-08-07 18:54:51			Anonymou				Yes
489	Neutral	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 18:50:54			Anonymou				Yes
490	Neutral	Cheap ' Wo	Blogs	http://infocheap.info/cheap-wood	infocheap.info	2013-08-07 18:49:15			Anonymou				Yes
491	Positive	Buy * Andr	Blogs	http://infocheap.info/buy-andrew-	infocheap.info	2013-08-07 18:48:57			Anonymou				Yes
492	Positive	Cheap Pric	Blogs	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 18:48:27			Anonymou				Yes
493	Neutral	Cheap ? Sn	Blogs	http://infocheap.info/cheap-smell	infocheap.info	2013-08-07 18:42			Anonymou				Yes
494	Neutral	Cheap Pric	Shopping	http://infocheap.info/cheap-price-	infocheap.info	2013-08-07 18:37:06			Anonymou				Yes
495	Neutral	Buy & Dogi	Blogs	http://infocheap.info/buy-dogit-fo	infocheap.info	2013-08-07 18:36:51			Anonymou				Yes
496	Neutral	Cheap – Gr	Blogs	http://infocheap.info/cheap-grena	infocheap.info	2013-08-07 18:36:18			Anonymou				Yes
497	Neutral	Offer " doT	Blogs	http://infocheap.info/offer-doterra	infocheap.info	2013-08-07 18:33:15			Anonymou				Yes
498	Neutral	Buy % Win	Blogs	http://infocheap.info/buy-winsom	infocheap.info	2013-08-07 18:31:15			Anonymou				Yes
499	Neutral	Cheap * W	Blogs	http://infocheap.info/cheap-wellr	infocheap.info	2013-08-07 18:27:15			Anonymou				Yes
500	Neutral	Cheap % W	Blogs	http://infocheap.info/cheap-winix-	infocheap.info	2013-08-07 18:26:27			Anonymou				Yes

This analysis will be incomplete without a treatment of the data analyzed by another tool to validate the output received by NetBase in a secondary way. This is covered in the next part, which are the results of processing the raw data indexed by NetBase in KH Coder, a statistical analysis tool. The aim is to see whether some fundamental attributes that are assumed to be taken care of by the data mining model used in NetBase holds true. A particular attribute in this regard is something like the co-occurrence of words.

4.2.KH Coder: Exploration of natural attributes of raw data indexed by NetBase

In this section, an examination is undertaken to validate some fundamental rules which ought to hold true in data analysis. For instance, there ought to be the co-occurrence of words in the English Language vocabulary or the hierarchical clustering of groups of words based on their associations in a text of the English Language. Other attributes of the NetBase data are also examined.

The KH Coder application is run on text files constructed from the 500 raw datasets indexed by NetBase. These represent the period of 12 months. This discussion will focus on just the attributes of the data itself.

Having introduced the co-occurrence analytical method in Chapter 3, it can be seen in Fig. 25 that correct associations are made between words that normally occur together. On further observation these associations of co-occurrence center on the Consumer Confidence Index. For example, the Consumer Confidence Index is usually published in the United States on Tuesday. As the output from KH Coder shows the words ‘released’ and ‘Tuesday’ (at the bottom of Fig. 25), co-occurred in the dataset contained the text file KH Coder analyzed.

The word ‘default’, is seen co-occurring with words that are known at the time to be associated with events of interest as well. For instance, there was a ‘shutdown’ of government services in the U.S due to an imminent default in honoring the debt obligations of the Federal Government at a time when the Republicans rejected the proposition made by the Obama administration to raise the borrowing limit. This co-occurrence shows up in the data

Other words that co-occurred and still do during conversations of economic matters are the British house price dilemma, the publication of revised consumer confidence index in the U.K.in July and the outlook of house prices and other sectors of the economy.

In Fig. 26, a non-functional attribute of the raw data set is examined. It shows the relationship between the frequencies of occurrence of terms (x-axis) with respect to the number of paragraphs (y-axis). The exponential relationship shown seems obvious, intuitively. Considering that a body of text or population of

raw data is about a particular subject of interest, it is inevitable that the volume of data's growth is a function of the number of the terms that it contains. That is the interpretation of the result shown in Fig. 26.

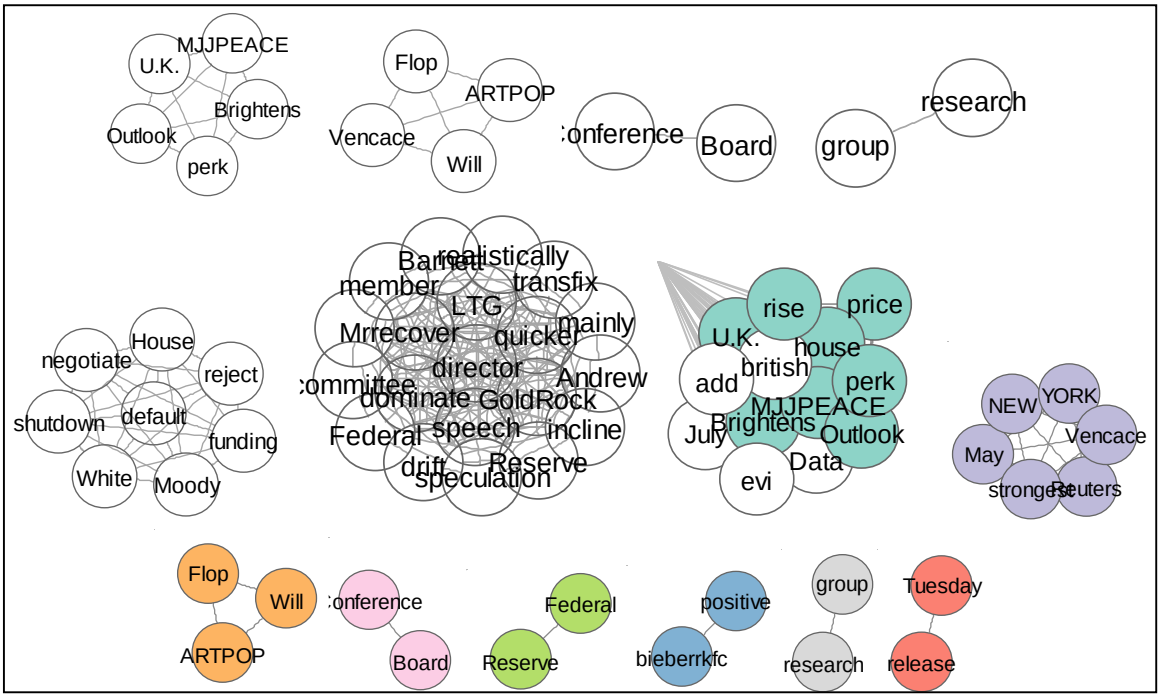


Fig. 25: Co-occurrence associations of indexed raw datasets

Fig. 26 highlights the fact that the terms contained in the raw dataset are relevant to the context of the keyword, albeit the terms sitting at around position 100 have a high density and by implication relevance to the context of the keyword. By extension they are relevant to the subject of the economic confidence of consumers.

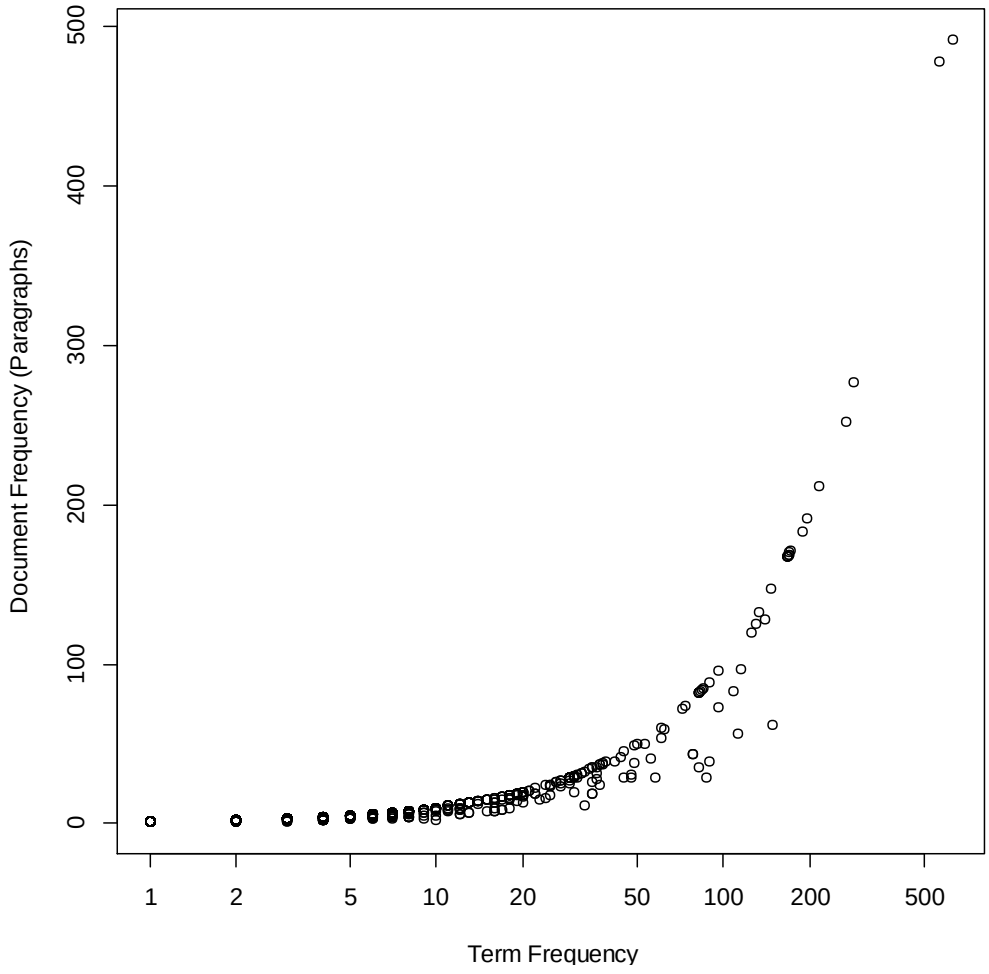


Fig. 26: Relationship between frequency of terms and the number of paragraphs in a set of data

Another angle from which the raw data set is examined is that of hierarchical clustering. Groups of associated words that are related to each other relative to the context of the subject matter are structured into a cluster. Among these clusters is a hierarchy of how they are used in the dataset. Their frequency of use and number of associations between clusters determines their position on the hierarchy of clusters. This is shown in Fig. 27.

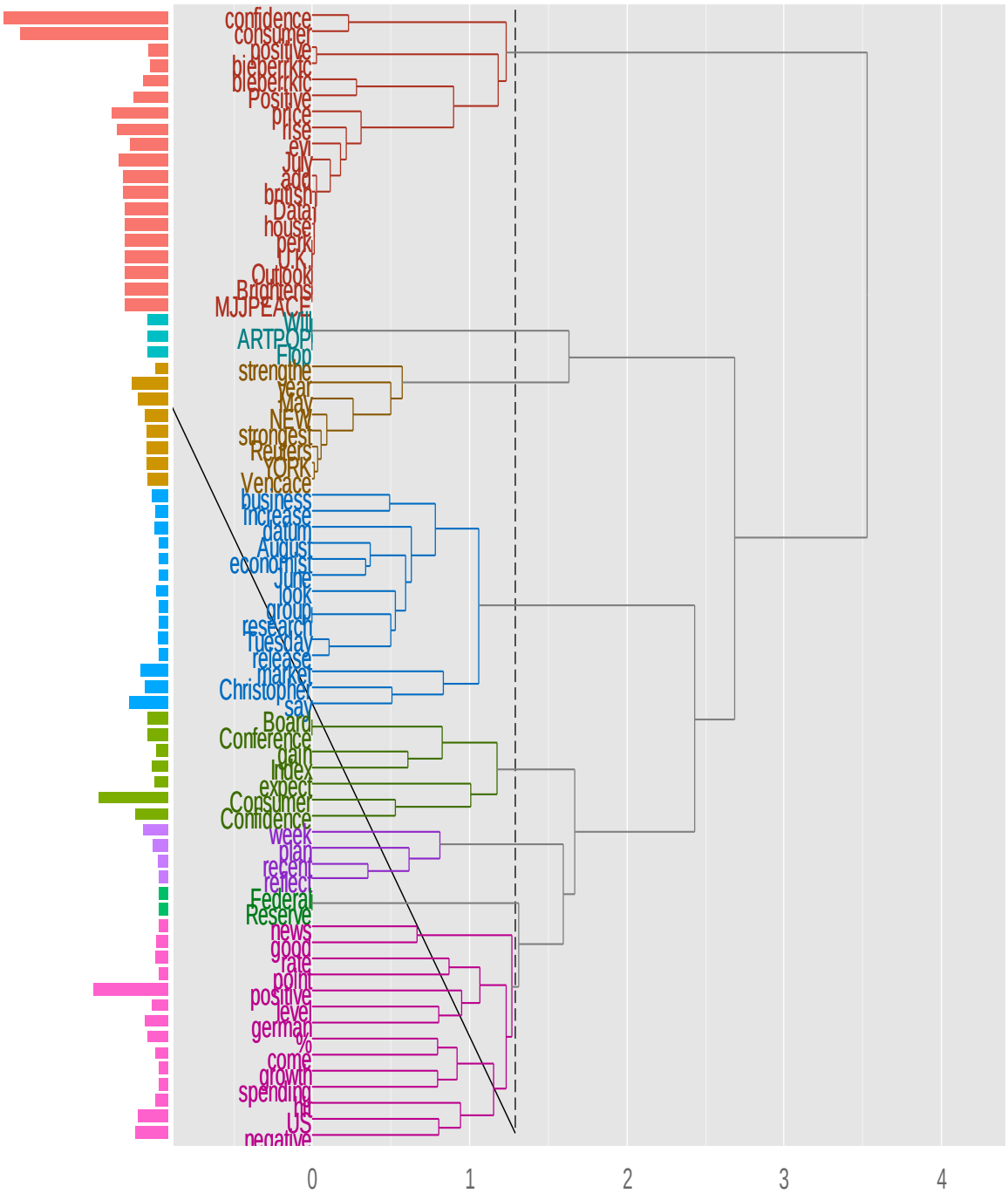


Fig. 27: Hierarchical cluster of terms contained in the raw data set indexed by NetBase®

The interpretation relative to the context of consumer confidence does give credence to the fact that terms such as: confidence, consumer, price, rise, data, house or outlook, feature a lot in the vocabulary used during formal and informal economic conversations. Whether it is on social media networks or in other forms of news, this shows the dataset is relevant to the subject of interest—consumer confidence.

Chapter 5

Summary & Conclusions

5.1. Summary

The purpose of this thesis has been an exploratory examination of the "sufficiency" of data from social media networks for the purposes of determining a higher order level of knowledge with a focus on factors related to the CCI.

A general introduction and history of this index was presented in Chapter 1, highlighting the need for its improvement with the advent of other sources of data such as social media networks. This approach opens the feasibility of taking advantage of the fact that obtaining data from social media is timely when compared to the traditional method that is used to derive the CCI. In this introductory Chapter, the objectives of the research were set and the research problem was stated—which is to evaluate whether data from social media networks are "good enough" to measure higher order levels of knowledge.

A review of the body of knowledge and ongoing research effort in the area of computational social science was reviewed. Chapter 2 presented an in-depth review of published literature in the area of data mining models and methodologies, where such concepts as NLP and LIWC were explored. It was established that there was a strong research interest to advance existing models for extracting data from social media networks for a variety of applications. These focused mainly on micro information synthesis, which falls short of being applicable for the synthesis of macro information.

In Chapter 3 the methodology for undertaking this research was presented alongside the data collection tool that was used—NetBase®. In this regard, fourteen keywords frequently used in formal and informal conversations about the economy and consumer confidence at large were identified. These were in turn used to search indexed records of conversations from across social media networks stored in NetBase's database.

With an outlay of how the research was executed, select keywords were used to gather data from NetBase® for periods of 3 months, 6 months and 12 months (October 2012 to October 2013). This was done by means of search, where NetBase® used the input keyword as a reference to extract conversations related to it. Each search term yielded three outputs from NetBase® (i) a graphical word cloud summary of attributes of a population of data related to the keyword (ii) a quantified tabular summary of attributes of a population of

data related to the keyword and (iii) 500 raw datasets from the data population related to the keyword used for the search.

These sets of data were presented in Chapter 4 and analyzed to elicit disadvantageous and advantageous characteristics of data from social media networks. For instance, in order to determine macro knowledge like the Consumer Confidence Index beyond the exploratory analysis, both qualitative and quantitative, KH Coder was used to examine some fundamental language rules to examine NetBase's data collection model through another lens.

The results from NetBase® were grouped into subcategories, including the Top terms, Top emotions expressed, most used hashtags, Top attributes, Top behaviors, Top brands and Top dis/likes. 'Top behaviors' was rejected on an analysis category because research showed that humans can better judge good behavior based on the context than a machine can. The same was for 'Top attributes'. Other subcategories were disregarded because they were peripheral to the subject of interest. These include Top brands, Top likes and Top dislikes.

As a consequence, only the Top emotions expressed, the Top terms and the most used hashtags were the subcategories examined throughout the analysis. And only seven keywords of the fourteen were analyzed in order to keep the content of the thesis concise.

The analysis in Chapter 4 shed light on the strengths and weaknesses of data from social media networks. Chief among its strengths was its reflection of topics of interest that are making the buzz at any one time. It is also real-time in nature and does not suffer any lags that hinder most forms of traditional data collection, e.g. CCSs. There is variety and volume when data from social media networks is considered as a source of information generation, which has been a shortcoming when other forms of data are put into some context of use beyond their primary domain.

Besides these advantageous attributes of data from social media networks, there are fundamental concerns around our understanding of the structure of this data. On a fundamental note, there is a knowledge gap in understanding the structure of this data. This in turn influences the accuracy of the data collection models which results in a high level of 'noisiness' in the data. Such noisiness could be traceable to model error and also lack of understanding of the nature of social media networks interactions to adjust other factors that could influence the accuracy of our data collection models. This is important when it has to do with accurately associating the right context to the intended meaning of a conversation especially with respect to a keyword of interest.

Other findings from the analysis in Chapter 4 also show that the hashtag which is used to make popular topics of interest easier to 'trend' seems irrelevant when it comes down to correctly determining higher levels of knowledge. Also, the weakness of the hashtag was further exposed, given the popular use of acronyms and slangs that have no connection with the English Language. In another area, the quantitative analysis of the results from NetBase® shows that in comparison to the figures being claimed as the number of monthly active users on a network like Facebook, that the number of conversations relevant to any particular chosen topic (like 'tax') can be relatively small.

Again this reinforces but diversifies the concern of the level of 'noisiness' of data from social media networks and calls into question their sufficiency for macro information synthesis. As was demonstrated, the highest top term in one of the results occurred in conversations had by only about 0.12 % of Facebook's monthly users alone. This is cause for concern, because it implies that either the data mining tool has a significant 'model' error or most conversations on social media networks are grossly irrelevant in the light of specific named topics of interest.

5.2. Conclusion

The CCI is an influential economic indicator relative to the indicative prospects of an economy. How it is measured is therefore vital to its continual success as an economic indicator in the near future, especially as the world's economy becomes strongly globalized.

Thus far, the dominant traditional method of CCI measurement has proven its reliability, but there is room for improvement owing to the number of times CCIs are revised after the first publication. An aspect of CCI measurement that is ripe for improvement is the area of its accuracy, the root cause of the multiple revisions.

To improve the accuracy of CCIs, it is logical to exploit other sources of data that would complement the traditional method of evaluating the CCS. This should include only different sources of data, but one whose data introduces a diversification of perspectives that would give a convergent view on the subject of interest—CCI. To this end, social media networks was identified as a potential new source of data, given their proliferation in the past decade and the number of [consumers] who are active users.

Social media networks not only contain data that is rich in variety, but they are a source of data that is readily accessible in real-time. This data can be extracted within seconds, minutes or hours, as desired. This is a distinctive advantage over the time it takes to perform a CCS.

CCI as a metric quantifies the subjective disposition of consumers about their economic prospects. For this reason data from social media networks that relate to this topic must be identified correctly, based on the potential information they give about the economic situation of users holding conversations. This process of identification and association is critical to the usefulness of any data extracted from social media networks to develop macro information like a CCI. The tool used to extract this data and the concept of data mining that such tool implements also affects the usefulness of such data.

The tool used to extract data for this research study from social media networks was NetBase®, a tool which implements the data mining concept of NLP. Keywords related to the economy, consumers and CCI in general were used to 'search' and 'extract' data.

After the data collection was performed, the three types of data collected were analyzed. This analysis qualitatively showed that data from social media networks has a strong correlation with popular topics of discussion in the period of time for which data was collected. It also revealed the challenges posed by slangs and acronyms when the prolific tagging system of hashtag is put under the spotlight. It showed that although hashtags are associated with making it is easy to identify popular topics that are trending, it does not always correlate with economic trends.

Furthermore, a quantitative analysis of the data showed that comparatively, the ratio of conversations deduced to be associated to a keyword (by NetBase®) and the total number of conversations had in the different social media platforms is not proportional. Consistently this number is small and it gives credence to the argument that data from social media networks is extremely noisy. It also calls into question whether data mining concepts such as NLP and their implicit implementation assumption(s) are suitable for the type of data that trails the unstructured nature of verbal conversations. Perhaps it is suggestive of a need for further research to focus on adapting these micro information synthesis concepts to serve macro information synthesis. There is also the need to analyze the increasing use of videos for which there exists little data mining concepts beyond specific concepts used by engineers to extract intangible information.

The verification analysis performed on the data using KH Coder does confirm that NetBase® and indeed the NLP concept correctly associate words that are related to a CCI keyword. Words that co-occur do in fact show a strong correlation in the Co-occurrence graph. The same goes for groups of words that relate to CCI being identified after a Cluster analytic method was performed in KH Coder. This gives confidence that at least the NLP concept is adequate for extracting data from social media networks to generate micro information.

In conclusion, this thesis has to some degree revealed that data from social media networks possesses the attributes of variety and being relevant in real-time. These are surely an advantage over the traditional method of CCS used to measure CCI. However, when explored further, the issue of correctly associating the context of conversations with the intentions of the participants arises, relative to the subject of interest at hand. Therefore, this would induce other forms of error that will make such data unreliable as a single source of measurement.

Overall, based on the analysis undertaken in this thesis, data from social media networks are "good enough", but as only **complementary** data for use with other forms of data to synthesize higher levels of knowledge. Therefore, rather than use data from social media networks alone, our advice is that other data formats with a set scope of context (surveys or interviews), are used as the source of primary data.

5.3. Future Study

This research study exposes numerous areas of work that could improve the future applicability of data from social media networks to the synthesis of macro information. Starting with the most important aspect of accuracy of association, concepts that are sufficient for micro information origination need to be developed further and tested extensively for use cases beyond a micro scope. Additional new concepts that are better suited to the unstructured form of data from social media networks needs to be developed.

Beyond the level of data association and extraction, methods that could be used to measure whether the data extracted is related to the keyword (and its context), should be developed. If they are related, there needs to be a method to measure how many continues to be unrelated and therefore represents the error of 'model rules' implemented in a given data mining tool.

Besides these fundamental areas for further research, it would be beneficial to look into emerging language forms that are unconventional. Examples discovered in this study includes hashtags and hash tagged acronyms in particular. It is perhaps a right time to seek to understand whether a new form of vocabulary is emerging, that would enable a better utilization of the insights they provide. It is believed this will alleviate some of the issues bordering on extraction of irrelevant data from irrelevant conversations.

Finally, it is recommended that a future effort should aim to extract larger data sets from social media networks, perhaps in the magnitude of thousands to millions. This should be done using a different data mining model to NLP, with a goal to assess whether the data gathered gives similar hints to the summaries deduced by NetBase® based on NLP. This would facilitate a comparison between the NLP and such model. Once this is achieved it will be more possible for an effort to be undertaken to actually derive an accurate index of Consumer Confidence.

Bibliography

Anon., n.d. *Office of Research Ethics and Integrity.* [Online]
Available at: http://www.research.uottawa.ca/ethics/consent.html
[Accessed 05 April 2015].

Asur, S. & Huberman, B. . A., 2010. *Predicting the Future With Social Media.* s.l., Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference. IEEE/WIC/ACM, pp. 492 - 499.

Bank of Ghana, 2011. *Bank of Ghana Monetary Policy Report- Real Sector Development.* [Online]
Available at:
http://www.bog.gov.gh/privatecontent/MPC_Press_Releases/real%20sector%20indicators%20of%20economic%20activity%20dec%2011%20-%20reviewedfinal.pdf
[Accessed 17 April 2015].

Barry, Chris; Markey, Rob; Almquist , Eric; Brahm, Chris, 2011. *Putting Social Media to Work.* [Online]
Available at: http://www.bain.com/Images/BAIN_BRIEF_Putting_social_media_to_work.pdf
[Accessed 17 April 2015].

BBC, 2014. *Boom and Bust.* [Online]
Available at: http://www.bbc.co.uk/bitesize/higher/history/usa/boombust/revision/1/
[Accessed 23 April 2015]

Berman, E. & Wang, X., 2011. *Essential statistics for public managers and policy analysts.* s.l.:CQ Press.

Berry, S. & Davey, M., 2004. How should we think about consumer confidence?. *Bank of England Quarterly Bulletin*, Autumn.

Blackshaw, P. & Nazzaro, M., 2004. Consumer-Generated Media (CGM) 101: Word-of-Mouth in the Age of the Web-Fortified Consumer.. *A Neilsen BuzzMetrics White Paper*, Spring.

Bloomberg News, April 2003. Consumer Confidence Shows a Substantial Gain. *The New York Times,* p. 8.

Bogleheads, 2013. *Financial Securities.* [Online]
Available at: http://www.bogleheads.org/wiki/Financial_securities
[Accessed 20 April 2015].

Casey, G. P. & Owen, A. L., 2013. Good News, Bad News, and Consumer Confidence. *Social Science Quarterly,* March, 94(1), p. 292–315.

Casteleyn, Jonathan, 2013. *Major divergence in consumer confidence signals declines for stocks.* [Online]
Available at: http://marketrealist.com/2013/02/major-divergence-in-consumer-confidence-signals-declines-for-stocks/
[Accessed 20 April 2015].

Chaffey, Dave, 2014. *Global social media research summary 2014: A compilation of social media statistics for consumer adoption and usage.* [Online]
Available at: http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
[Accessed 20 April 2015].

Charles Arthur, The Guardian, 2014. *Facebook emotion study breached ethical guidelines, researchers say.* [Online]
Available at: http://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say
[Accessed 21 April 2015].

Chitwan, R., 2002. *The Life History Calendar as a Data Collection Tool.* [Online]
Available at: http://perl.psc.isr.umich.edu/Res_Brief_2_final.pdf
[Accessed April 2015].

Collobert, R. et al., 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research,* Volume 12, pp. 2461-2505.

Cotsomitis, J. A. & Kwan, A. C. C., 2006. Can Consumer Confidence Forecast Household Spending? Evidence from the European Commission Business and Consumer Surveys. *Southern Economic Journal,* January, 72(3), pp. 597-610.
Curtin, R. T., 2002. *Surveys of Consumers: Theory, Methods, and Interpretation.* Washington DC, s.n.

Daas, P. J. & Puts, M. J., 2014. *Social Media Sentiment and Consumer Confidence.* [Online]
Available at: https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf
[Accessed April 2015].

Dan Noyes - Zephora Internet Marketing Solutions, 2014. *The Top 20 Valuable Facebook Statistics - Updated October 2014.* [Online]
Available at: https://zephoria.com/social-media/top-15-valuable-facebook-statistics/
[Accessed 12 April 2015].

Dominitz, J. & Manski, C. F., 2004. How Should We Measure Consumer Confidence?". *Journal of Economic Perspectives,* Spring, 18(2), pp. 51-66.

Elizabeth Dwoskin, The wall Street Journal , 2013. *Twitter's Data Business Proves Lucrative.* [Online]
Available at: http://online.wsj.com/news/articles/SB10001424052702304441404579118531954483974
[Accessed 12 April 2015].

Elson, S. B. et al., 2012. *Using Social Media to guage Iranian Public Opininon and Mood After the 2009 Election.* [Online]
Available at: http://www.rand.org/content/dam/rand/pubs/technical_reports/2012/RAND_TR1161.pdf
[Accessed April 2015].

Encyclopædia Britannica, 2015. *Social status.* [Online]
Available at: http://www.britannica.com/EBchecked/topic/551450/social-status
[Accessed 07 April 2015].

Erik Qualman, 2009. *Statistics Show Social Media Is Bigger Than You Think.* [Online]
Available at: http://www.socialnomics.net/2009/08/11/statistics-show-social-media-is-bigger-than-you-think/
[Accessed 03 April 2015]

Fredrickson, B. L. & Losada, M. F., 2005. Positive Affect and the Complex Dynamics of Human Flourishing. *American Psychologist,* October, 60(7), pp. 678-686.

Freedman, D. et al., 1988. The Life History Calendar: A Technique for Collecting Retrospective Data. *American Sociological Association,* Volume 18, pp. 37-68.

FxPro Blog, 2014. *U.S. Michigan consumer sentiment index revised down to 88.8.* [Online]
Available at: http://blog.fxpro.co.uk/u-s-michigan-consumer-sentiment-index-revised-down-to-88-8/
[Accessed 20 April 2015].

Garner, A. C., 1981. Economic determinants of consumer sentiment. *Journal of Business Research,* 9(2), p. 205–220.

Gillin, P. & Schwartzman, E., 2010. Social marketing to the business customer: Listen to your B2B Market, generate major account leads, and build client relationships. In: *Social marketing to the business customer.* illustrated ed. s.l.:John Wiley & Sons, p. 256.

GNIP, 2010. *How Does Netbase Achieve the Best Accuracy for Understanding Consumers Online?.* [Online]
Available at: http://gnip.com/docs/Listening_Accuracy_White_Paper.pdf
[Accessed 14 April 2015].

Gritten, A., 2011. New insights into consumer confidence in financial services. *International Journal of Bank Marketing,* 29(2), pp. 90 - 106.

Higuchi, K., 2013. *Cluster Analysis: Indicate Clusters by Different Colors.* [Online]
Available at: http://khc.sourceforge.net/en/gallery/pages/image/imagepage6.html
[Accessed 07 April 2015].

Higuchi, K., 2013. *Co-occurrence Network: Betweenness Centrality.* [Online]
Available at: http://khc.sourceforge.net/en/gallery/pages/image/imagepage10.html
[Accessed 18 April 2015].

Higuchi, K., 2013. *Correspondence Analysis.* [Online]
Available at: http://khc.sourceforge.net/en/gallery/pages/image/imagepage15.html
[Accessed 07 April 2015].

Higuchi, K., 2014. *KH Coder.* [Online]
Available at: http://khc.sourceforge.net/en/
[Accessed 07 April 2015]

Hogenboom, A. et al., 2010. *Mining Economic Sentiment using Argumentation Structures.* s.l., In Advances in Conceptual Modeling- Applications and Challenges. Springer Berlin Heidelberg, pp. 200-209.

Hopkins, D. J. & King, G., 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science,* 54(1), pp. 229-247.

Hudszinki, D., 2014. *A private discussion on Nielsen's market intelligence product for retail: Consumer Confidence monitoring* [Interview] [12 December 2014].

Hulbert, Mark, 2010. *It's darkest before the dawn- Commentary: Bull markets often begin before consumer confidence hits bottom.* [Online]
Available at: http://www.marketwatch.com/story/contrarian-view-on-consumer-confidence-2010-02-24?reflink=MW_news_stmp
[Accessed 21 April 2015].

Hymans, S. H., Ackley , G. & Juster, T. F., 1970. Consumer Durable Spending: Explanation and Prediction. *Brookings Papers on Economic Activity,* Volume 2, pp. 173-206.
IBM, 2014. *What is Watson? As a cognitive technology, Watson is a natural extension of what humans can do at their best.* [Online]
Available at: http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html
[Accessed 20 April 2015].

INSEAD, 2009. *How Obama used social networking tools to win.* [Online]
Available at: http://knowledge.insead.edu/innovation/how-obama-used-social-networking-tools-to-win-1600
[Accessed 17 April 2015].

INSEE- National Institute of Statistics and Economic Studies, 2004. *Monthly consumer confidence survey.* [Online]
Available at: http://www.insee.fr/en/themes/indicateur.asp?id=20
[Accessed 02 April 2015].

Investopedia, 2013. *Present Situation Index.* [Online]
Available at: http://www.investopedia.com/terms/p/presentsituationindex.asp
[Accessed 14 April 2015].

Investopedia LLC, 2015. *Complete Guide to Corporate Finance.* [Online]
Available at: http://www.investopedia.com/walkthrough/corporate-finance/2/taxes/types-taxes.aspx
[Accessed 18 April 2015].

Investopedia, LLC, 2015. *Capital Expenditure - CAPEX.* [Online]
Available at: http://www.investopedia.com/terms/c/capitalexpenditure.asp
[Accessed 14 April 2015].

Investopedia, 2013. *Expectations Index.* [Online]
Available at: http://www.investopedia.com/terms/e/expectationsindex.asp
[Accessed 14 April 2015].

Investopedia, 2014. *Capital Investment.* [Online]
Available at: http://www.investopedia.com/terms/c/capital-investment.asp
[Accessed 14 April 2015].

Jansen, W. & Nahuis, N. J., 2003. The stock market and consumer confidence: European evidence. *Elsevier,* April, Issue 1, pp. 89-98.

Jeremy M. Piger, 2003. *Consumer Confidence Surveys: Do They Boost Forecasters' Confidence?.* [Online]
Available at: https://www.stlouisfed.org/publications/re/articles/?id=415
[Accessed 11 April 2015].

Jim Edwards - Business Insider, 2013. *Twitter's 'Dark Pool': IPO Doesn't Mention 651 Million Users Who Abandoned Twitter.* [Online]
Available at: http://www.businessinsider.com/twitter-total-registered-users-v-monthly-active-users-2013-11?IR=T
[Accessed 19 April 2015].

Jurafsky, D. & Manning, C., 2014. *Natural Language Processing.* [Online]
Available at: https://class.coursera.org/nlp/lecture/124
[Accessed 06 April 2015].

Karen Christensen - Forbes India, 2014. *Incidental Vs. Integral: Understanding your Emotions.* [Online]
Available at: http://forbesindia.com/article/rotman/incidental-vs.-integral-understanding-your-emotions/36949/1
[Accessed 11 April 2015].

Katona, G., 1960. *The powerful consumer,* s.l.: McGraw-Hill.
Keller Fay Group, 2010. *Four WOM Statistics.* [Online]
Available at: http://www.kellerfay.com/four-wom-statistics/
[Accessed 3 April 2015].

Kongthon , A., Sangkeettrakarn , C., Kongyoung , S. & Haruechaiyasak , C., 2009. *Implementing an online help desk system based on conversational agent.* New York, Proceedings of the International Conference on Management of Emergent Digital EcoSystems. ACM, p. 69.
Kramer, D., Guillory, J. & Hancock, J., 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 25 03, pp. 8788-8790.

Kwan, A. C. C. & Cotsomitis, J. A., 2006. The Usefulness of Consumer Confidence in Forecasting Household Spending in Canada: A National and Regional Analysis. *Economic Inquiry,* 44(1), pp. 185-197.

Lemmens, A., Croux, C. & Dekimpe, M. G., 2007. Consumer Confidence in Europe: United in diversity?. *International Journal of Research in Marketing,* June, 24(2), pp. 113-127.

Liu, B., 2010. Sentiment Analysis: A Multi-Faceted Problem. *IEEE Intelligent Systems,* 25(3), pp. 76-80.

Lovell, M. C., 1975. Why was the Consumer Feeling So Bad?. *Brookings Papers on Economic Activity,* Volume 2, pp. 473-479.

Ludvigson, S. C., 2004. Cosumer Confidence and Consumer Spending. *Journal of Economic Perspectives,* Spring, Volume 18, pp. 29-50.

Madigan, Kathleen ;, 2015. *U.S. Consumer Confidence Surges to Highest Level Since 2007.* [Online]
Available at: http://www.wsj.com/articles/u-s-consumer-confidence-surges-to-highest-level-since-2007-1422371844
[Accessed 17 April 2015].

Mangold, G. W. & Faulds, D. J., 2009. Social media: The new hybrid element of the promotion mix. *Elsevier,* 52(4), p. 357–365.

MasterCard Worldwide, 2010. *MasterCard Worldwide Index of Consumer Confidence.* [Online]
Available at:
http://www.masterintelligence.com/view_report/index_of_consumer_confidence/africa/h1_2010/Nigeria
[Accessed 14 April 2015].

McKinsey & Company, 2012. *The social economy: Unlocking value and productivity through social technologies.* [Online]
Available at: http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_social_economy
[Accessed 11 April 2015].

Merkle, D. M., Langer, G. E. & Sussman, D., 2004. *Consumer Confidence: Measurement and Meaning.* New York, NY, ABC News.

Microsoft Research, 2014. *Natural Language Processing.* [Online]
Available at: http://research.microsoft.com/en-us/groups/nlp/
[Accessed 12 April 2015].

Mishne, G. & Glance, N., 2006. Predicting Movie Sales from Blogger Sentiments. *American Association for Artificial Intelligence*, Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 155-158.

Mowen, J. C., 1999. *The 3M Model of Motivation and Personality-Theory and Emperical Applications to Consumer Behaviour.* Massachusetts: Kluwer Academic Publishers.

Mueller, E., 1963. Ten Years of Consumer Attitude Surveys: Their Forecasting Record. *Journal of the American Statistical Association,* December, Volume 58, pp. 899-917.

Mueller, E., 1963. Ten Years of Consumer Attitude Surveys: Their Forecasting Record. *Journal of the American Statistical Association,* 58(304), pp. 899-917.

Naoi, T., Yamada, T., Iijima, S. & Kumazawa, T., 2011. Applying the caption evaluation method to studies of visitors' evaluation of historical districts. *Tourism Management,* Volume 32, pp. 1061-1074.

National Restaurant Association, 2014. *State of the American Consumer.* [Online]
Available at: http://www.restaurant.org/News-Research/News/State-of-the-American-consumer
[Accessed 18 April 2015].

Netbase Solutions, 2012. *Social Intelligence Warehouse Drives Instant Answers and Action.* [Online]
Available at: http://www.netbase.com/social-intelligence/consumer-base/
[Accessed 02 April 2015].

Netbase, 2010. How Does Netbase Achieve the Best Accuracy for Understanding Consumers Online?. *Netbase*, September.
Available at: https://gnip.com/docs/Listening_Accuracy_White_Paper.pdf
Accessed 23 April 2015

Netbase, 2012. [Online]
Available at: http://www.netbase.com/
[Accessed 01 April 2015].
NetBase, 2014. *Why Use NetBase?.* [Online]
Available at: http://www.netbase.com/social-media-management/why-use-netbase/
[Accessed 01 April 2015].

Nielsen, 2013. *Consumer COnfidence Concerns and Spending Intentions around the World.* [Online]
Available at: http://www.nielsen.com/us/en/insights/reports/2014/consumer-confidence-concerns-and-spending-intentions-around-the-world.html
[Accessed 11 April 2015].

Nielsen, 2013. *Global Consumer Confidence.* [Online]
Available at: http://www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/global-consumer-confidence.html
[Accessed 11 April 2015].

Nielson, 2013. *Consumer COnfidence Concerns and Spending Intentions around the World.* [Online]
Available at: http://www.nielsen.com/us/en/insights/reports/2014/consumer-confidence-concerns-and-spending-intentions-around-the-world.html
[Accessed 11 April 2015].

Nielson, 2014. *Global Consumer Confidence.* [Online]
Available at: http://www.nielsen.com/us/en/solutions/measurement/global-consumer-confidence.html
[Accessed 11 April 2015].

Nigam, K. & Hurst, M., 2004. *Towards a Robust Metric of Opinion. In AAAI Spring Symposium on Exploring Attitude & Affect in Text.* [Online]
Available at: http://www.kamalnigam.com/papers/metric-EAAT04.pdf
[Accessed 08 April 2015].

Nyu.edu, 2014. *Consumer Confidence.* [Online]
Available at: http://pages.stern.nyu.edu/~nroubini/bci/ConsumerConfidence.html
[Accessed 08 April 2015].

O'Connor, B., Balasubramanyan, R., Routledge, B. R. & Smith, N. A., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM,* Volume 11, pp. 122-129.

Oest, R. v. & Franses, H. P., 2008. Measuring changes in consumer confidence. *Journal of Economic Psychology,* June, 29(3), pp. 255-275.

Office of National Statistics, 2013. *Social Networking: The UK as a Leader in Europe,* United Kingdom: Office of National Statistics. [Online]
Available at: http://www.ons.gov.uk/ons/rel/rdit2/internet-access---households-and-individuals/social-networking--the-uk-as-a-leader-in-europe/sty-social-networking-2012.html
[Accessed 18 April 2015]

Olowofeso, O. E. & Doguwa, S., 2012. *Consumer Sentiment and Confidence Indices in Nigeria: A Panel Data Analysis.* [Online]
Available at: http://www.bis.org/ifc/events/6ifcconf/olowofesodoguwa.pdf
[Accessed 18 April 2015].

Oppewal, H., Paas, L. J., Crouch, G. I. & Huybers, T., 2010. Segmenting consumer based on how they spend a tax rebate: An analysis of the Australian stimulus payment. *Journal of Economic Psychology,* August, 31(4), pp. 510-519.
Otoo, M. W., 1999. Consumer Confidence and the Stock Market. *Board of Governors of the Federal Reserve System Research Paper Series - FEDS Papers,* November.pp. 99-60 .

Out:Think, n.d. *The 6 types of Social Media.* [Online]
Available at: http://outthinkgroup.com/tips/the-6-types-of-social-media
[Accessed 01 April 2015].

Oxford Dictionaries, 2015. *Oxford Dictionaries - Language matters.* [Online]
Available at: http://www.oxforddictionaries.com/
[Accessed 06 April 2015].

Pak, A. & Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC,* May, Volume 10, pp. 1320-1326.

Pennebaker, J. W., Booth, R. J. & Francis, M. E., 2007. *Linguistic Inquiry and Word Count: LIWC2007.* [Online]
Available at: http://psycholingua.googlecode.com/svn-history/r79/trunk/resources/LIWC2007_OperatorManual.pdf
[Accessed 20 April 2015].

Performics, 2010. *Social Networking Study: Facebook Use Continues to Rise; Brand Participation and Engagement Heavily Welcomed by Social Networkers.* [Online]

Available at: http://www.performics.com/social-networking-study-facebook-use-continues-to-rise-brand-participation-and-engagement-heavily-welcomed-by-social-networkers/
[Accessed 12 April 2015].

PR Newswire, 2013. *The Conference Board's Consumer Confidence Index rebounds in April,* s.l.: PR Newswire Association LLC.

Research at Google, Google Inc., 2014. *Natural Language Processing.* [Online]
Available at: http://research.google.com/pubs/NaturalLanguageProcessing.html
[Accessed 19 April 2015].

Reuters, 2014. *U.S. consumer confidence falls in November to lowest since June.* [Online]
Available at: http://www.reuters.com/article/2014/11/25/us-usa-economy-confidence-idUSKCN0J91O920141125
[Accessed 06 April 2015].

Roberts, I. & Simon, J., 2001. *What Do Sentiment Surveys Measure?.* [Online]
Available at: http://www.rba.gov.au/publications/rdp/2001/pdf/rdp2001-09.pdf
[Accessed 20 April 2015].

Rouse, Margaret, 2011. *Capex (capital expenditure).* [Online]
Available at: http://whatis.techtarget.com/definition/CAPEX-capital-expenditure
[Accessed 11 April 2015].

Rouse, Margaret, 2014. *Social Media.* [Online]
Available at: http://whatis.techtarget.com/definition/social-media
[Accessed 11 April 2015].

Ryan Barnes, Investopedia, 2009. *Economic Indicators: Consumer Confidence Index (CCI).* [Online]
Available at: http://www.investopedia.com/university/releases/consumerconfidence.asp
[Accessed 02 April 2015].

Ryan, J. - Bloomberg, 2014. *U.K. Consumer Confidence Stalls as Economic Outlook Deteriorates.* [Online]
Available at: http://www.bloomberg.com/news/2014-11-28/u-k-consumer-confidence-stalls-as-economic-outlook-deteriorates.html
[Accessed 06 April 2015].

Schweidel, D. A., Moe, W. W. & Boudreaux, C., 2012. Social Media Intelligence: Measuring Brand Sentiment from Online Conversations. *Marketing Science Institute*, pp. 12-100.
Scribd., 2015. *About - The word's favourite open publishing platform.* [Online]
Available at: https://www.scribd.com/about
[Accessed 11 April 2015].

Sergie, M.A.; Kayakiran, F.; Bloomberg, 2015. *Qatar, Shell Scrap $6.5 Billion Project After Oil's Drop.* [Online]
Available at: http://www.bloomberg.com/news/2015-01-14/qatar-shell-scrap-6-5-billion-project-amid-

oil-price-collapse.html
[Accessed 16 April 2015].

Sr. Bogue, Edith;, 2008. *Correlation and Regression SPSS.* [Online]
Available at: http://www.slideshare.net/edithosb/correlation-and-regression-spss
[Accessed 20 April 2015].

Statista, 2014. *Statistics and Market Data on Social Media & User-Generated Content.* [Online]
Available at: http://www.statista.com/markets/424/topic/540/social-media-user-generated-content/
[Accessed 20 April 2015].
Tausczik, Y. R. & Pennebaker, J. W., 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology,* 29(1), p. 24–54.

The Conference Board of Canada, 2014. *Consumer Confidence.* [Online]
Available at: http://www.conferenceboard.ca/topics/economics/consumer_confidence.aspx
[Accessed 06 April 2015].

The Conference Board of Canada, 2014. *Consumer Confidence US.* [Online]
Available at: http://www.conferenceboard.ca/topics/economics/consumer-confidence-us.aspx
[Accessed 06 April 2015].

The Conference Board, 2011. *Consumer Confidence Survey Technical Note - February 2011.* [Online]
Available at: https://www.conference-board.org/pdf_free/press/TechnicalPDF_4134_1298367128.pdf
[Accessed 06 April 2015]

The Conference Board, 2011. *Consumer Confidence Survey® Technical Note – February 2011.* [Online]
Available at: http://www.conference-board.org/pdf_free/press/TechnicalPDF_4134_1298367128.pdf
[Accessed 06 April 2015].

The Conference Board, 2014. *The Conference Board Consumer Confidence Index® Declines.* [Online]
Available at: http://www.conference-board.org/press/pressdetail.cfm?pressid=5397
[Accessed 12 April 2015].

The New York State Department of Taxation and Finance, 2014. *More Tax Types.* [Online]
Available at: http://www.tax.ny.gov/forms/more_tax_types.htm
[Accessed 28 April 2015].

The Next Web, Inc., 2013. *Facebook passes 1.23 billion monthly active users, 945 million mobile users, and 757 million daily users.* [Online]
Available at: http://thenextweb.com/facebook/2014/01/29/facebook-passes-1-23-billion-monthly-active-users-945-million-mobile-users-757-million-daily-users/
[Accessed 21 April 2015].

The Nielsen Company, 2014. *Nielson: Global Consumer Confidence Holds Steady to Close 2013; While Spending Intentions Retreat to Q4 2012 Levels.* [Online]
Available at: http://www.nielsen.com/us/en/press-room/2014/nielsen-global-consumer-confidence-

holds-steady-to-close-2013.html
[Accessed 04 April 2015].

The Wall Street Journal, 2013. *The Oracle of Cyberspace?- The Online Buzz about Warren Buffet joining Twitter.* [Online]
Available at:
http://online.wsj.com/article/SB10001424127887324766604578459392169116764.html?mod=WSJ_Books_LS_Books_5
[Accessed 25 April 2015].

The Wall Street Journal, 2014. *November U.S. Consumer Confidence Unexpectedly Falls to 88.7.* [Online]
Available at: http://online.wsj.com/articles/november-u-s-consumer-confidence-unexpectedly-falls-to-88-7-1416928693
[Accessed 25 April 2015].

Thelwall, M., Buckley, K. & Paltoglou, G., 2011. Sentiment in Twitter Events. *Journal of the American Society for Information Science and Technology,* 62(2), p. 406–418.

Thomson Reuters & University of Michigan, 2014. *Surveys of consumers.* [Online]
Available at: http://data.sca.isr.umich.edu/fetchdoc.php?docid=50688
[Accessed 14 April 2015].

Tobin, L., 2012. *Entrepreneur: How to Start an Online Business.* 1st ed. Chichester: Wiley & Sons.

TradingEconomics, 2014. *United Kingdom Consumer Confidence.* [Online]
Available at: http://www.tradingeconomics.com/united-kingdom/consumer-confidence
[Accessed 28 April 2015]

Tsytsarau, M. & Palpanas, . T., 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery,* 24(3), pp. 478-514.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M., 2010. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* s.l., ICWSM, pp. 178-185.

Twitter Inc., 2014. *Character Counting.* [Online]
Available at: https://dev.twitter.com/overview/api/counting-characters
[Accessed 21 April 2015].

Twitter, Inc., 2014. *Using hashtags on Twitter.* [Online]
Available at: https://support.twitter.com/articles/49309-using-hashtags-on-twitter#
[Accessed 21 April 2015].

unesco.org, 2008. *Correspondence Analysis.* [Online]
Available at: http://www.unesco.org/webworld/idams/advguide/Chapt6_5.htm
[Accessed 7 April 2015].

United Nations, 2015. *World Economic Situation and Prospects 2015.* [Online]
Available at: http://www.un.org/en/development/desa/publications/wesp-2015-2.html
[Accessed 3 April 2015].

Vanguardngr, 2011. *MasterCard ranks Nigeria most optimistic market in Africa.* [Online]
Available at: http://www.vanguardngr.com/2011/08/mastercard-ranks-nigeria-most-optimistic-market-in-africa/
[Accessed 3 April 2015]

Weisstein, Eric W. - From MathWorld -- A Wolfram Web Resource, 2015. *Sampling Theorem.* [Online]
Available at: http://mathworld.wolfram.com/SamplingTheorem.html
[Accessed 12 April 2015].