TSALLIS ENTROPY & BIG DATA



SUBMITTED

BY

WAGISH KUMAR RAI

DIVISION OF PHYSICS & APPLIED PHYSICS

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

A final year project report

presented to

Nanyang Technological University

in partial fulfilment of the

requirements for the

Bachelor of Science (Hons) in Physics

Nanyang Technological University

30 April 2016

Abstract

The analysis of data has been of increasing interest in the recent decade because of the rapid progress in technology, rising amount of useful data being produced as well as many potential problems it may address such as global warming, financial crashes, ecological transitions. [1]

Traditionally, computer science has the most impact on this area of research because it employs effective methods such as machine learning, neural-networks and more recently deep-learning to make predictions from data. In fact, they are most rigorously used in the industry as well as research. They are very stable tools in-fact and are not too much on the bleeding edge. This however is one main category of tools available to a researcher or an individual if he or she is interested in working on problems relating to big data analytics. [1]

The desire for a universal indicator that may aid in predictive analytics is very traditional from an epistemological point of view, such that one may observe that in other fields, researchers seek certain universal ideas. This epistemology is not free from criticisms, however. [2]

An argument as to why Tsallis entropy is a viable candidate to be explored in big data will be explored for and argued based on 1.) Validity and reliability of Tsallis Entropy 2.) Tsallis entropy in past applications 3.) my own Tsallis Entropy application and possible conclusions from the data-sets I have worked on exploring a novel application of Tsallis index negative q as an indicator.

Table Of Contents

Abstract i

Table of Contents ii

Chapter 1 Introduction

1.1 Big Data

1.2 Research In Big Data 1

1.3 Entropy 1

1.4 Generalized Entropy Information Measures

Chapter 2 Tsallis Entropy & Applications

2.1 2

2.2 2

Chapter 3 Application of Tsallis Entropy on Financial Tweets Data

3.1 Twitter Dataset Analysis Methodology 3

3.2 Twitter Dataset Analysis 3

Chapter 4 Conclusions and Future Work…………………………………………….... 4

4.1 Conclusions 4

4.2 Recommendatin in Future Work

References ………………………………………………………………………………. 5

Chapter 1: Introduction

1.1 Big Data

The subject of Big Data has grown in both importance and popularity because of the vast number of problems it presents as well possible innovations in academia and industry. Below are some of the main reasons.

1 Technological Progress

The fast development of computers and computer power has increased significantly. Computing power that was once only accessible to large companies are now available to a wider community.The price drop of computing hardware due to innovations in the industry has allowed the renting of computing power and space instead of requiring people to invest large amounts of money. For example, anyone can rent Amazon Elastic Compute Cloud (Amazon EC2) to speed up computing drastically. [1]

2 Exponential Growth in Data

Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, remote sensing, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2016, every day 2.5 exabyte (2.5×1018) of data are created. [1]

3 Value of Big Data

These data are used for predictions that interest businesses, governments and academia as analysis of data can lead to better decision making, operational efficiency and reduced risk.

1.2 Research in Big Data

From the premise of the rising importance of data, most research conducted tend to multi-disciplinary and the skill sets required to tackle problems or innovate are diverse and deep. The most prominent field involved is computer science. Computer scientists tend to use methods such as machine learning and more recently deep learning to answer data science questions. There is also heavy work done by the statistics department as well as Physicists and mathematicians who share similar expertise. Some use statistical mechanics ideas to approach problems, however are marked with controversy as not all the data are produced systematically from first principles. Reproducibility and constructing models on systems with no first-principles backing is actively debated in the philosophy of science community and also produce problems when the models do not universally fit in all given circumstances. As such, fields like complexity science emerge but do not necessarily seem address issues on reproducibility and non-first principle problems directly with great epistemological consistency and clarity.

First principle is a basic, foundational proposition or assumption that cannot be deduced from any other proposition or assumption. In physics and other sciences, theoretical work is said to be from first principles, or *ab initio*, if it starts directly at the level of established science and does not make assumptions such as empirical model and fitting parameters. For example, calculation of electronic structure using Schrödinger's equation within a set of approximations that do not include fitting the model to experimental data is an *ab initio* approach. [9]

There is no stopping of scientists from working solely on first-principle related problems but it happens that most of the problems humanity is facing tend to come in the form on non-first principled problems. Examples include global warming, earthquakes, terrorism and financial crashes. As such, it may not be reasonable to divert from such significant problems as solving or alleviating them in any helpful manner has to be carried out and attempted more for its potential pragmatic humanitarian benefits instead of the purpose of fulfilling academic consistency.

In light of the significant criticisms presented and the importance of problems presented, it may be substantial to look at this subject from an epistemological perspective of Feyerabend’s exploring tools such as Tsallis entropy that may help to approach such problems.

Since the data are generated by non-first principle based procedures, we may attempt to find certain index or measure of the data, characterize the data and draw some possible conclusions. As argued by the philosopher of science Paul Feyerabend that there are no necessarily exclusive fixed rules in substantiating all of the progress of science and knowledge [5].

Introduction – Entropy

One such tool that we may want to look at is entropy. It may be used as a detection tool of which above or below certain value for a system (data) may lead to certain predictions.

Often, the system has to have certain fixed properties and axioms to align itself with in the picture of a particular form of entropy.

Some complications include on whether the system (or the way data is produced) resides with first principles, non – first principles or a mix should not be a starting major concern, although in normal circumstances it would be from my opinion. As that would influence the type and parameters of system we would be considering.

As such, a candidate tool might be some mathematical structure of entropy that can take a wide range of system as such not to be obsessed with the problems devised by the possible strict definitions of a system but instead is able to encompass a wider range of systems, such as the generalized Tsallis entropy for example.

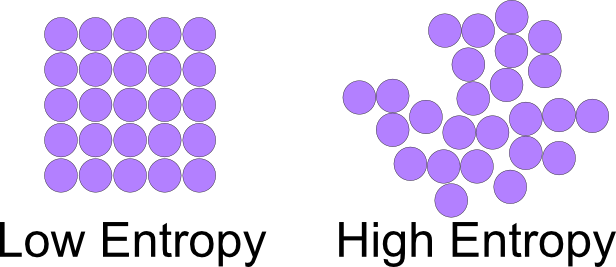


Fig.1 Higher Order for Lower Entropy and Higher Disorder for Higher Entropy

One common, and general qualitative understanding of entropy we may have is that lower values of entropy associates with order and higher values of entropy associates itself with disorder. It may seem that this understanding of entropy may be superficial and insignificant to an expert working in the field, however this understanding is sufficient to employ its application on data-sets.

The mathematics behind this simple qualitative understanding is the following:

On what the Tsallis entropy values implies for that data-set or application seems to be a diverse subject and tends to vary from different fields [4][5][6]. They do not have any universal meaning, but they have universal and consistent mathematical structures to rely on entropy that make them useful tools for various applications. And the equations for different entropy types varies, some are more generalized to take in more possible cases, axioms or modifications on axioms (fig.2).

In general, I would treat entropy as a measure of order and disorder of a probabilistic variable in a data-set. In other words, entropy of a data-set is simply the measure of the amount of order and disorder of that variable in the data-set.

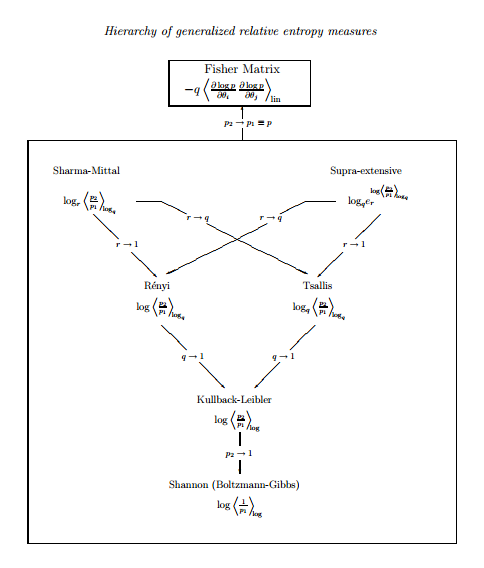


Fig.2 Generalizations of entropy [10]

Examples of these generalizations is expressed in fig.2. Historically, Boltzmann-Gibbs entropy was the first landmark work on statistical mechanics. The project did not stop at Boltzmann-Gibbs entropy alone, because the assumptions and axioms seem to be limited and one could use a little bit of imagination and reasonable speculation that there may be systems beyond what the BG entropy can cover. In fact, it is well known that BG entropy only covers well for short-range interactions. Long-range interactions such as gravity cannot be expressed or understood from a statistical mechanics perspective. This is only one of many physical motivations to come up with other generalized forms of entropy, there also have been non-physical motivations to come up with other entropies from information theory and other fields as well, such as cybernetics. The one on formulating Tsallis entropy came from a purely physics or physical motivation [3].

In the case of long-range interactions, it deals with the subject matter called non-extensive statistical mechanics that includes Tsallis entropy. As there is no clear picture on how the additional descriptive property such as long-range interactions of Tsallis entropy may benefit us in studying large data-sets. Hence, one possible methodology we might employ is to do a comparative study of the data-set to that of BG entropy.

Chapter 2: Tsallis Entropy & Past Applications

Many people – notably the US physicist Murray Gell-Mann, who won the 1969 Nobel Prize for Physics for his theoretical work on elementary particles – agree that Tsallis entropy is a true generalization of Boltzmann–Gibbs entropy. But there are many detractors too, among whom the principal charge is that the Tsallis index q is a mere “fitting parameter” for systems that are not well enough understood. Naturally, Tsallis disagrees. If the fitting-parameter accusation were true, he says, it would not be possible to obtain q from first principles – as he did in 2008, together with quantum physicist Filippo Caruso, who was then at the Scuola Normale Superiore di Pisa in Italy. Tsallis and Caruso showed that q could be calculated from first principles for part of a long, 1D chain of particle spins in a transverse magnetic field at absolute zero. The value of q, which was not equal to one, reflected the fact that quantum effects forced some of the spins to form strong correlations (Phys. Rev. E 78 021102).

This calculation required a knowledge of the exact microscopic dynamics, which is not, however, always possible. In situations where the dynamics are not known, says Tsallis, then q indeed has to be obtained from fitting experimental data, but he claims that doing so is no different to how other accepted theories are employed in practice. As an example, Tsallis cites the orbit of Mars, which could be calculated from first principles – but only if both the distribution of all the other planets at a given moment, and the initial conditions of masses and velocities, were all known. Clearly, he says, that is impossible. “For the specific orbit, astronomers collect a lot of data with their telescopes, and then fit that data with the elliptic form that comes out of Newton’s law [of gravitation], and then you have the specific orbit of Mars,” he adds. “Well, here, it’s totally analogous. In principle, we would always like to be able to calculate q purely from mechanics, but it’s very hard, so q often has to be obtained from fitting.”

Chapter 3 Application of Tsallis Entropy on Financial Tweets Data

Methodology

The one I will work on applying BG entropy, Tsallis entropy and on a Twitter historical Data-Set that relates to finance related tweets during the 2007 financial crisis.

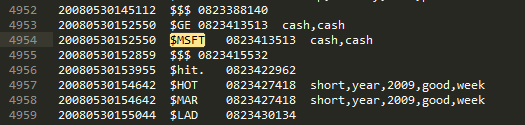


Fig.3 Part of the raw data-set

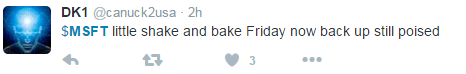


Fig. 4 an example of a tweet with $MSFT tag

The left-most column represents the time-stamp of the tweet given in yyyy/mm/dd/hh/mm/ss format. The second Colum represents the tag of the tweet messages (fig.4). The next column is the tweet ID and the last column is filtered tweet message that was used for sentiment analysis.

As observed, there are many financial related tags but some tags are not identifiable to relate to any specific financial entity. One way I approached this problem was to clean the data for Dow Jones 30 companies and analyze the Dow Jones 30 stock market crash with respect to the tweet data given here.



Fig.5 List of Dow 30 Symbols

Therefore, in our application we would be concerned with adding the $ symbol on the left of Dow 30 Symbol(s) and filtering data for tags for example $AAPL, $GS, $NKE etc.

After filtering the data, we would like to have a few plots to analyze the data. The first basic plot we can have is a monthly frequency plot. Basically, a measure of the number of Dow 30 Symbols tag mentions in a month.

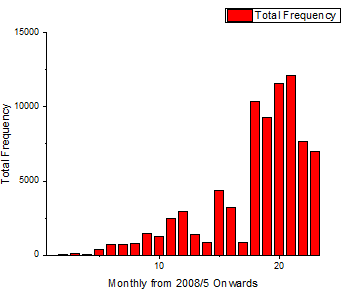


Fig.6 Dow 30 symbols total frequency/month plot shows a general increasing trend in the number increase of tweets starting from May 2008 to end 2009

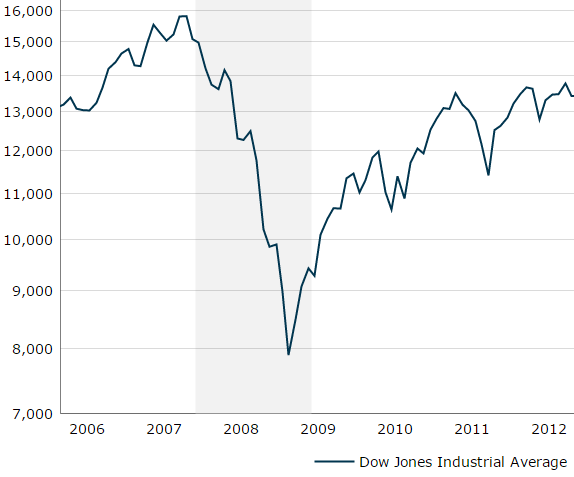


Fig.7 Dow Jones industrial average, grey area representing financial crisis period []

Looking at fig. 6 and fig. 7 we are able to confirm that the number of tweets increased during the crash 2008 – 2009. We may infer that there was an increase in interest in Dow 30 stocks during the recovery of Dow 30 stocks. After the recovery, interest started to drop gradually after 2010 in the Twitter Space deducing from the gradual drop of frequency of tweets of Dow 30 tags.

This prior analysis shows that the data may be useful for further analysis. One proposed way we can look at it is to study the entropy of the data and observe if anything meaningful emerges.

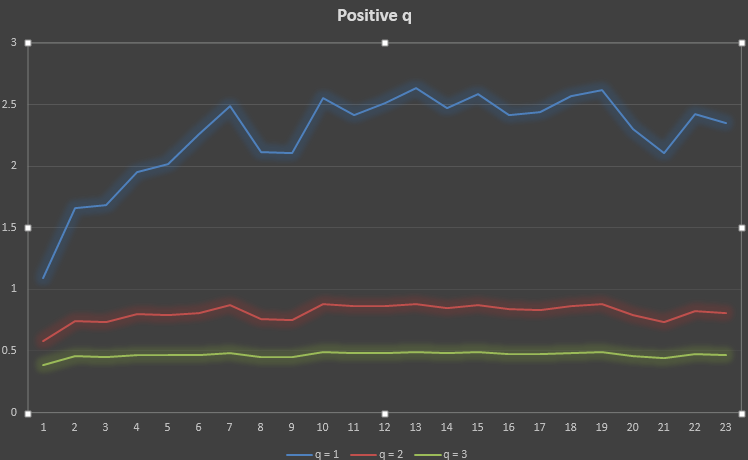
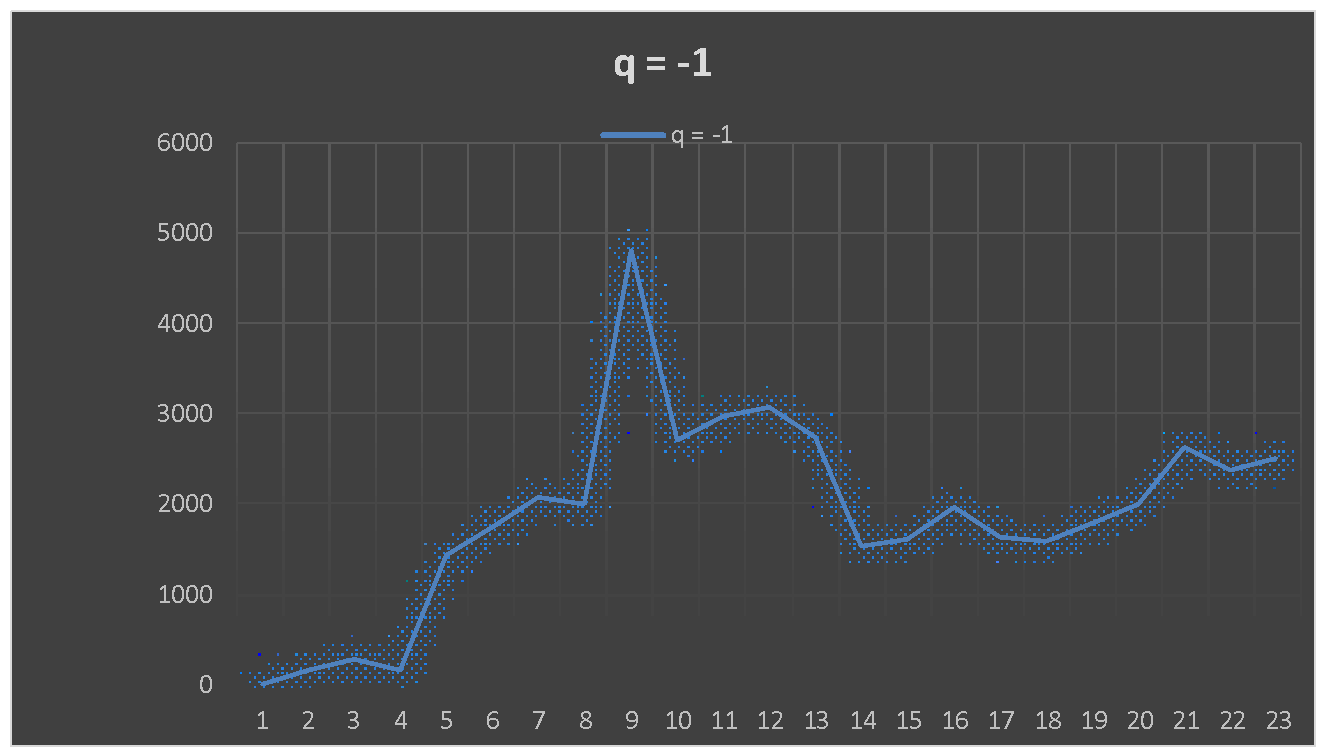
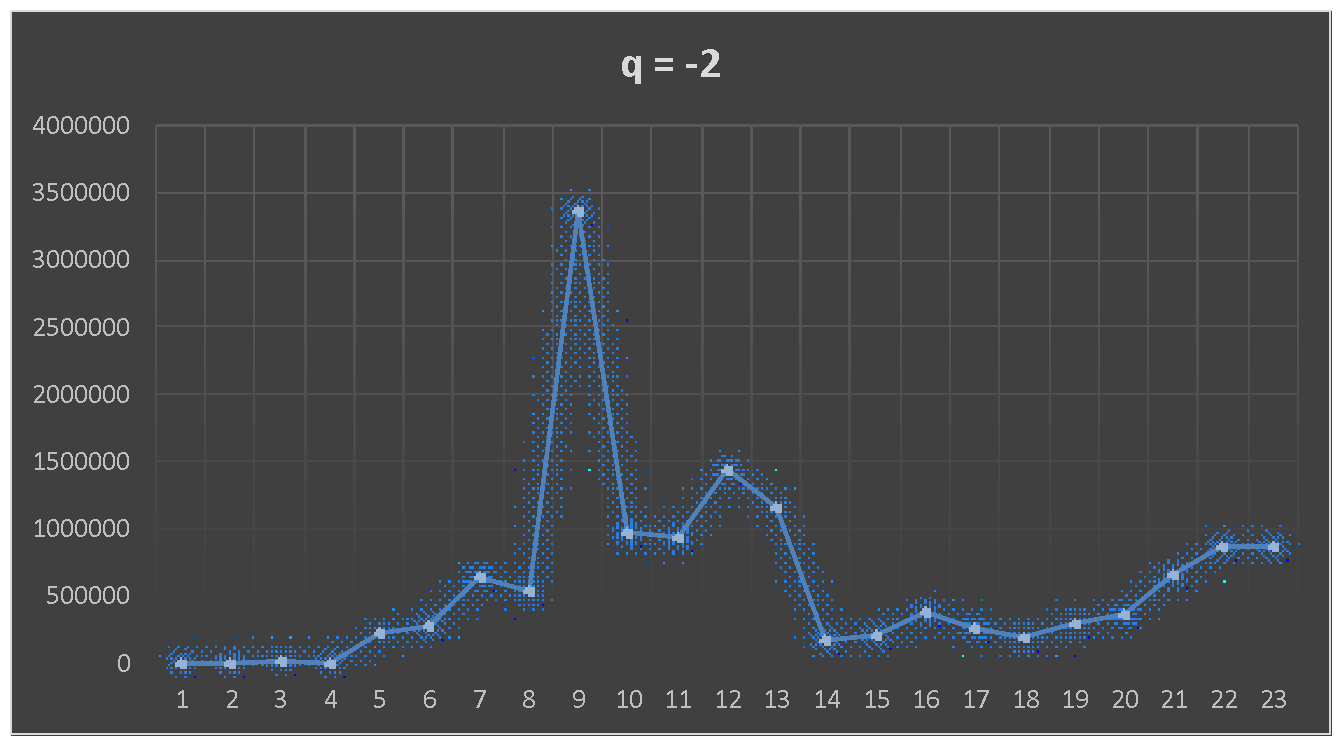
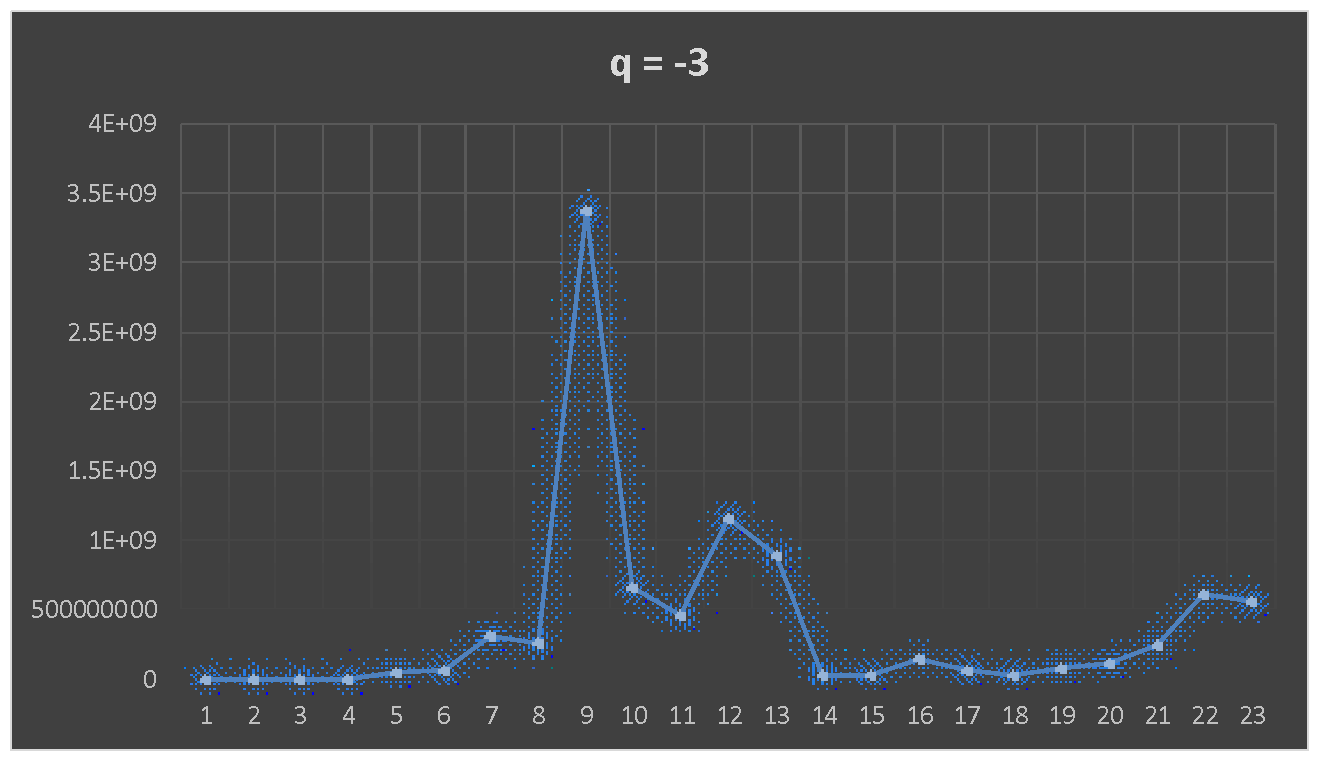


Fig.8 Tsallis Entropy Measures for positive values of q. For q = 1, we have the standard BG entropy plot of the tweets

Fig. 10 Tsallis Entropy for q = -1

Fig. 11 Tsallis Entropy for q = -2

Tsallis Entropy for q = -3

It is observed that for negative values of q for Tsallis entropy, peaks are more well defined and the Tsallis entropy values scales up significantly.

The peaks for negative q of Tsallis entropy do not correspond to any trend on the frequency of dow 30 tweets (monthly).

This suggests that certain properties of twitter data relating to dow 30 using negative q of for Tsallis entropy may be studied.

Bibliography

1. [Big data - Wikipedia, the free encyclopedia. Available at:](http://paperpile.com/b/jO40ir/fs7b) <https://en.wikipedia.org/wiki/Big_data.> [(Accessed: 21st April 2016)](http://paperpile.com/b/jO40ir/fs7b)

3. Marco Masi, Generalized information-entropy measures and Fisher information, arxiv (2008)

4. M. Portes de Albuquerque, I.A. Esquef, A.R. Gesualdi Mello, Image thresholding

using Tsallis entropy, Pattern Recognition Letters, (2004)

5. A.R Plastino, A. Plastino, Stellar polytropes and Tsallis' entropy,

Physics Letters A,(1993)

6. E.M.F. Curado and C. Tsallis, Generalized statistical mechanics: connection with

thermodynamics , J. Phys. A 24, L69-L72 (1991); Corrigenda 24, 3187 (1991); 25,

1019 (1992).