Automated Financial Text Analysis from Securities and Exchange Commission Filings

Darshil Shah[1], Parthiv Patel[2]

[1]*Student of Information Technology Dept, Thadomal Shahani Engineering College, Mumbai, India*
[2]*Student of Electronics Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*

---***---

Abstract - *Artificial Intelligence has become a very useful tool for the corporate world to focus at the fundamentals of the business systems which are related to manufacturing, entertainment, medicine, marketing, engineering, finance and other services. This helps the firms make in their regular reporting practices more efficiently. Dealing with complex and multiplex situations can be done very easily and immediately with these automated systems thus saving a lot of time. Changes to the language and construction of financial reports have strong implications for firms' future returns. These help in constructing frameworks and algorithms that have helped the business systems in achieving the optimum results. With information such as customer reviews, opinions and sentiment regarding products and services now readily available online, business have realised the immense potential of the insights inherent in this data and have started to use analytical concepts such as sentiment analysis subsumed as social media analytics to extract the desired information. These reporting changes are concentrated in the management discussion section. Thus sentimental analysis not only helps firms take huge decisions but also helps in knowing everything about any business challenges well in advance.*

Key Words: **Sentiment analysis, data extraction, financial decisions**

1. INTRODUCTION

The web, as a worldwide system of interconnection, provides a link between billions of devices and other people round the world. The fast development of social networks causes the tremendous growth of users and digital content. It opens opportunities for individuals with varied skills and information to share their experiences and knowledge with one another. There are several websites like Yelp, Wikipedia, Flickr, etc. that use the facility of the web to assist their users build optimum selections.

Sentiment analysis is the use of natural language for processing analysis of text, computational linguistics, and interpretation and classification of emotions (neutral, positive and negative) within text data using text analysis techniques. Sentiment analysis tools permit businesses to spot client sentiment toward product, brands or services in on-line feedback. Some money establishments have begun investment in departments that specialize in AI and machine learning applications that would confirm their customer's sentiments towards market developments.

2. CURRENT MARKET SITUATION

We researched the current market to see where AI comes into play within the finance business and to answer the subsequent questions:

1) How is sentiment analysis presently utilized in the finance world?

2) What results are the reason for sentiment analysis being so famous in finance?

This report covers vendors providing code across 3 applications:

•Data Search and Discovery

•Report Generation

•Process Automation

The paper is structured as follows. Section 2 describes the dataset. Section 3 presents the accuracy of classification for a wide range of text processing methods and machine learning algorithms. Section 4 explores the performance analysis. Section 5 concludes.

3. RELATED WORK

Sentiment Analysis has been heavily employed by businesses for social media opinion mining, particularly within the industry, wherever customers' feedback area unit is important. Within the recent year, it's been gaining quality within the finance sector, it's been used to analyze tweets of important monetary analysts and call manufacturers. It appears like there area units are most potential to be unbarred for the usage of Sentiment Analysis. It's curious to what is going to be a consequent breakthrough!

In 2013 Lazaridou et al. tried to learn meanings of a phrase by mistreatment integrative spacing linguistics models. In 2013 Chrupala used an easy repeated network (SRN) to find out continuous vector representations for sequences of characters. They used their model to unravel a personality level text segmentation and labeling task. A meaningful search area via Deep Learning is often created by mistreatment of repeated Neural Networks. Socher et al. in 2011, used algorithmic autoencoders for predicting sentiment distribution and planned a semi-supervised approach model. In 2012 Socher et al. proposed a model for linguistics integratively with the flexibility to find out compositional vector illustration for sentences of capricious length. Their planned model may be a matrix-vector algorithmic neural network model, algorithmic Neural Tensor Network (RNTN) design planned. RNTN use word vector and a analyse tree to represent a phrase and so use a tensor-based composition operate to calculate vectors for higher nodes.

Current approaches to mine sentiments from money texts mostly think about domain specific dictionaries. However, lexicon based strategies typically fail to accurately predict the polarity of economic texts. This paper aims to enhance the progressive and introduces a completely unique sentiment analysis approach that employs the conception of economic and non-financial performance indicators. It presents associate association rule mining based mostly graded sentiment classifier models to predict the polarity of economic texts as positive, neutral or negative. The performance of the planned model is evaluated on a benchmark money dataset. The model is additionally compared against different progressive lexicon and machine learning based mostly approaches and also the results area unit found to be quite promising. The novel use of performance indicators for money sentiment analysis offers attention-grabbing and helpful insights.

4. DATASET

We extracted the data from SEC / EDGAR financial reports. This website contained all the financial terms that were required. Each text file consists of at least one to three subsections. We divided the subsections into three different variables namely "Management and Discussion Analysis", "Qualitative and Quantitative Risk Analysis" and "Risk Factors". All these variables contained some of the positive, negative, constrain and uncertain terms. Humans Understand the meaning of emotions and they can segregate the terms associated with it. We created a list of words having all the positive, negative, certain and uncertain terms. While extracting it is easy for the test cell to get trained beforehand.

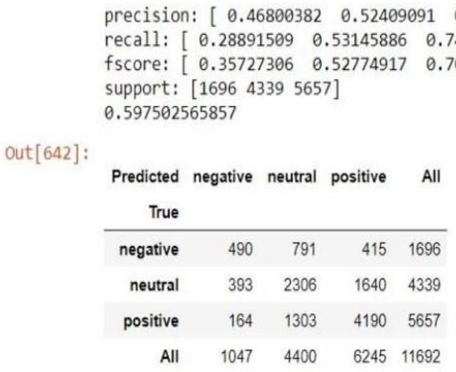
We extracted 153 text forms from the website and 42 different financial variables containing different terms. Each row contained a financial text form and its different variable counts. Since not all text forms contained the mentioned variables, we had given a null value to the missing terms. It is worthwhile to consider that each variable had different amounts of terms which was easy for us to analyse.

5. EXTRACTION METHODOLOGY

We extracted the data using python library "Beautiful Soup." It is an open source library used for pulling data out of HTML and XML files. It works with your favourite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forums.

6. TEXT EXTRACTION:

The World Wide Web has a cluster of text information in the form of useful text. This text is used for text analysis which further transforms it into more than one interpretation. This text and information remains isolated unless and until it automates the different types of discovery techniques. We can extract text from only those websites which contain tables of information. Systems like NLP models, Clustering Models can extract information from the web page having high reliability on the HTML tags

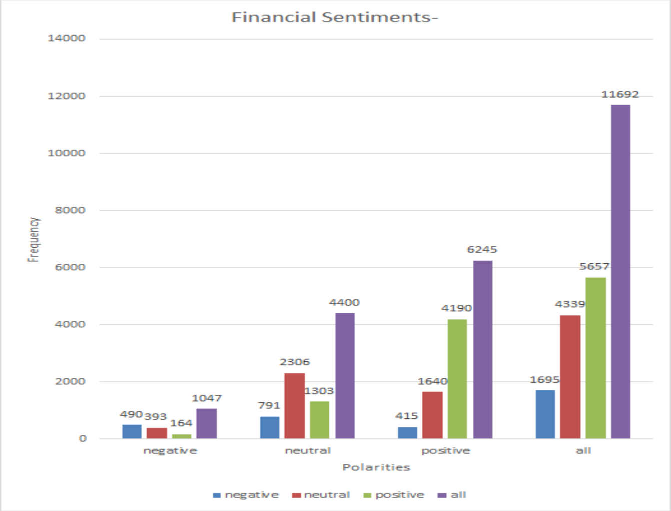


```
precision: [ 0.46800382  0.52409091  0.67093675]
recall: [ 0.28891509  0.53145886  0.74067527]
fscore: [ 0.35727306  0.52774917  0.70408335]
support: [1696 4339 5657]
0.597502565857
```

Out[642]:

Predicted	negative	neutral	positive	All
True				
negative	490	791	415	1696
neutral	393	2306	1640	4339
positive	164	1303	4190	5657
All	1047	4400	6245	11692

The system can extract information based on a high accuracy. This is based on all the tags used for fixing the table entries. However such systems cannot handle a large number of data text sets which are in the form of a narration. Unfortunately, such systems cannot handle the large proportion of text data that is in narrative form. There is lots of progress being done on many machine learning algorithms and NLP for extracting the systems. All the techniques required need to be in a logical as well as in a grammatical sentences. Often it is found on the web that the sentences are in the form of a summary and not the usual grammatical sentences where the systems cannot predict what to extract and with what algorithms. A new word technique for web pages is used for demonstrating NLP techniques which can be used for extracting the data in the form of non-grammatical text found on the web. The information depends on the individual relationship and facts. A single isolated fact is not at all enough to identify the primary key word for search engine optimization. The domain used in this paper is a financial sentimental text web page, extracting all the financial terms from the text pages. This financial text is associated with the companies having all the financial transactions, statements, analogies, etc. The output is presented in the form of “Positive” words, “Negative” words, “Neutral” words, “Constraint” words. Further we have calculated the polarity, frequency and the nature of the text. A typical NLP information extraction system parses each sentence, then applies rules based on the syntactic relation of phrases within a sentence. Such a system will find no useful syntactic clues in the text. Worse yet, a system that treats each phrase ending with a period as a separate sentence will have difficulty associating "CHANCE OF TRADE" with "LONG" or "SHORT" trading strategies. The text page source takes the text as its input and applies its algorithmic rules based on its layout. This further divides the text into rules of a formal argument and reasonably divides those text that are passed into the NLP systems. Webfoot handles a wide range of web page styles, including pages whose layout is indicated by HTML tags or by blank lines and white space, and pages with information in tabular or narrative format. This greatly expands the range of text data that can be extracted automatically from web pages. The NLP system used in these experiments is “Beautiful Soup”, which learns domain-specific text extraction rules from the previous data sets and examples. The rest of this paper describes the Web pages extraction and anomalies systems and presents empirical results for the domain of financial sentimental analysis. The combination of NLTK and NLP achieve surprisingly good performance for a system operating without the aid of syntactic knowledge. This opens the way for automatic analysis of a class of text data that has been largely inaccessible.



7. CONCLUSIONS

In this paper, we present an effective and robust general-purpose text analysis and extraction algorithm, which can automatically detect and extract text from complex background images. Our main future work involves using a suitable existing OCR and Anomalies technique that are used for recognizing the extracted text. The contributions of the proposed method are:

(a) Can handle both html and normal text web pages.

(b) Not sensitive to any robust with respect to font, sizes, orientations, alignment, uneven illumination, perspective and reflection effects

(c) It can be used for calculating all the trade and financial terms which can further be used for building financial strategies.

REFERENCES

[1] 1. Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. “Sarcasm as contrast between a positive sentiment and negative situation”, In EMNLP 2013, pp. 704–714.

[2] 2. Subrahmanian, V. S. and Diego Reforgiato. “Ava: Adjective-verb-adverb combi-nations for sentiment analysis”, In Intelligent Systems, 23(4):43–50. 2008.

[3] 3. B. Agarwal, N. Mittal, “Prominent Feature Extraction for Review Analysis: An Empirical Study”, In Journal of Experimental and theoretical Artificial Intelligence, 2014, DOI: 10.1080/0952813X.2014.977830.

[4] 4. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, “Lexicon-based methods for sentiment analysis”, Computational Linguistics, v.37 n.2, p.267-307, 2011

[5] 5. Esuli A., Sebastiani F. “SentiWordNet: A publicly available lexical resource for opinion mining”. In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), pages 417–422, 2006.

[6] 6. P Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of 40th Meeting of the Association for Computational Linguistics, pages 417–424, Philadelphia, PA.

[7] 7. http://eci.nic.in/eci_main1/GE2014/PC_WISE_TURNOUT.htm

[8] 8. Mariana Romanyshyn(2013). Rule-Based Sentiment Analysis of Ukrainian Reviews. International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, July 2013

[9] 9. S Bandyopadhyay and K Mallick, "A New Path Based Hybrid Measure for Gene Ontology Similarity", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.11, no. 1, pp. 116-127, Jan.-Feb. 2014, doi:10.1109/TCBB.2013.149

[10] 10. N. Mittal, B. Agarwal, S. Agarwal, S. Agarwal, P. Gupta, "A Hybrid Approach for Twitter Sentiment Analysis", In 10th International Conference on Natural Language Processing (ICON) 2013. pp.116-120.

[11] 11. A. Bakliwal, J. Foster, J. V. D. Puil, R. O"Brien, L. Tounsi, M. Hughes, "Senti-ment analysis of political tweets: Towards an accurate classifier". In Proceedings of NAACL Workshop on Language Analysis in Social Media, pages 49–58, 2011

[12] 12. Di Caro, L., & Grella, M. (2012). Sentiment analysis via dependency parsing. Computer Standards & Interfaces.

[13] 13. K.. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heil-man, D. Yogatama, J. Flanigan, N.A. Smith. "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments", In Proceedings of ACL, 2011.

[14] Xiaoqing Liu and Jagath Samarabandu, "Multiscale Edge-Based Text Extraction from Complex Images" in Content-Based Access of Image and Video Libraries, 1999. (CBAIVL '99), 1999, Proceedings. IEEE Workshop on, pp. 109–113.

BIOGRAPHIES

DARSHIL SHAH, Student of Information Technology Department, Thadomal Shahani Engineering College, Mumbai 400104

Parthiv Patel, Student of Electronics Department, Dwarkadas J Sanghvi College of Engineering, Mumbai 400104