TEXT MINING FINANCE NEWS OF FOUR BIG AMERICAN COMPANIES

R. Srinithy¹, V. Karpagam¹, J. Pamina²

¹PG Students, Sri Krishna College of Technology, Coimbatore, Tamil Nadu
²Assistant Professor, Sri Krishna College of Technology, Coimbatore, Tamil Nadu

ABSTRACT

Perception from textual database refers generally to the process of withdrawing interesting or non-retrieval patterns or knowledge from unorganized text documents. This is also called as text data mining or knowledge discovery. Due to huge amount of unstructured text data generated on the Internet today, text mining is believed to have high mercantile value. It can be viewed as a supplement of data mining or knowledge from organized databases. At the current scenario, the stored information is getting bigger or getting enlarged day by day. This is an unorganized form so we can not extricate the needed information. Some data mining techniques are used to extricate the useful information from the text documents, such as classification, clustering, visualization and information extraction. Classifications of text mining work as text categorization, text clustering, association rule extraction and trend analysis are done based on applications. This paper introduces the research status of text mining of financial news of certain companies. Here framework of text mining with techniques is discussed as well as benefits and limitations of text mining have been discussed and concluded that text mining is an advanced and sensational concept that controls overloaded information problem by using certain techniques.

I. INTRODUCTION

Text mining in financial information also portrays a great impact in today’s scenario. Financial information mainly lies in financial statements, but it is to be observed that financial information can be textual as well as numerical. There are many origins for data like annotation, demonstrable reports, managerial letters etc, which are different from the financial statements that provide useful valuable information for decision makers. As a form of communication, text data allow researchers to understand managers’ behavioral perspectives and business behavior. The evaluation of these data requires text mining methods. Text mining is a large data analysis used in the evaluation of semi-structural and non-structural data. Within the scope of this information, it is aimed to raise consciousness on the use of text mining in non-structural data evaluation in the field of financial information. Hence the study is on the process of text mining the financial news of four big American companies, Starbucks, Kraft, Wal-Mart and Mondelez through the concept of data
II. TOKENIZATION

Tokenization is the act of breaking up a concatenation of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are disposed. The tokens become the input for processes like parsing and text mining. Tokenization is used in computer science, where it plays a large part in the process of lexical analysis. A token is a meaningful unit of text, most often a word, that we are interested in using for further analysis, and tokenization is the process of splitting text into tokens. This concept uses “unnest_tokens” to break text into individual tokens and transform it to a tidy data, that is one-row-per-term-per-document. There are various concepts under tokenization, that include,

Basic tokenizers is the concept where the functions perform basic tokenization into words, sentences, paragraphs, lines, and characters. The functions can be piped into one another to create at most two levels of tokenization. For instance, one might split a text into paragraphs and then word tokens, or into sentences and then word tokens. A character vector or a list of character vectors to be tokenized into n-grams. If x is a character vector, it can be of any length, and each element will be tokenized separately. If x is a list of character vectors, each element of the list should have a length of 1.

 Chunk text into smaller segments, is a concept where a text or vector/list of texts, are given that breaks the text into smaller segments each with the same number of words. This allows you to treat a very long document, such as a novel, as a set of smaller documents. For instance, A character vector or a list of character vectors to be tokenized into n-grams. If x is a character vector, it can be of any length, and each element will be chunked separately. If x is a list of character vectors, each element of the list should have a length of 1.

 Count words, sentences, characters, are taken as input texts. These functions use the string i package, so they handle the counting of Unicode strings (e.g., characters with diacritical marks) in a way that makes sense to people counting characters.

 N-gram tokenizers, these functions tokenize their inputs into different kinds of n-grams. The input can be a character vector of any length, or a list of character vectors where each character vector in the list has a length of 1. For instance, a character vector or a list of character vectors to be tokenized into n-grams. If x is a character vector, it can be of any length, and each element will be tokenized separately. If x is a list of character vectors, each element of the list should have a length of 1.

 III. VISUALIZATION OF HIGHEST tf-idf WORDS

The statistic of tf_idf is intended to measure how important or necessary a word is to a document for example, to one novel in a collection of novels or to one website to a collection of websites. The idea of tf-idf is to find the important words for the content of each document by decreasing the weight for commonly used words
and increasing the weight for words that are not used very much in a collection or corpus of documents. Calculating tf-idf attempts to find the words that are important or common in a text, but not too common.

HIGHEST TF_IDF WORDS FOR EACH COMPANY

Visualization of top terms for each company

IV. SENTIMENT ANALYSIS

When human readers approach a text, we use our understanding or knowledge of the emotional intent of words to infer whether a section or part of text is positive or negative, or perhaps characterized by some other more emotions like surprise or disgust. We can use the tools of text mining to approach the emotional content of text programmatically.

The three general-purpose lexicons are

- AFINN from Finn Årup Nielsen,
- bing from Bing Liu and collaborators, and
- nrc from Saif Mohammad and Peter Turney.

All three of these lexicons are based on unigrams, i.e., single words. These lexicons contain many English words and the words are assigned scores for positive/negative sentiment, and also possibly emotions like joy, anger, sadness, and so forth. The nrc lexicon categorizes words in a binary fashion into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise,
and trust. The Bing lexicon categorizes words in a binary fashion into positive and negative categories. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

**AFINN SCORE**

The Loughran dictionary divides words into six sentiments: “positive”, “negative”, “litigious”, “uncertainty”, “constraining”, and “superfluous”.

**V. CONCLUSION**

Text mining is an evolving field in research and many different computer programs are needed to be developed to serve various requirements of its applications. These computer programs will help to evaluate the textual data sources conveniently and will enable researchers to identify new horizons of its usefulness. Financial information has an important function in providing communication between the internal and external stakeholders of businesses. Financial information mainly lies within the boundaries of the financial statements, but it should be noted that financial information can be textual as well as numerical. The factors such as changing environmental conditions and technology have deeply influenced the conceptual basis phenomena in the definition of assets and the needs of the stakeholders. Along with the technology, changes in the asset structure of the enterprises have occurred and the intangible assets have gained weight rather than the tangible assets. Hence various concepts
under tokenizers has been analyzed and sentiment analysis concept has also been portrayed in detail. The paper provides a keen understanding on tokenization, visualization and analysis of sentiments of the four concerned companies under the concept of data science using R.

VI. REFERENCES


