IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

A Study on Text Analytics and Categorization Techniques for Text Documents

S. Swathi[1], P. Lalitha[2]

Assistant Professor in Department of MCA in Chaitanya Bharathi Institute of Technology, Hyderabad, India[1,2]

Abstract: Text Mining is termed as extraction of relevant yet hidden information from the text document. One of the essential concepts in the field of text mining is Text classification. Through the sudden growth in digital world and available documents, the task of organizing text data becomes one of the principal problems. The order issue need been broadly examined clinched alongside information mining, machine learning, database, and majority of the data recovery. On the foundation of quick majority of the data processing, we need made An investigation of backing vector machine in quick classification. Eventually Tom's perusing presenting that essential rule from claiming SVMs, we portrayed those transform from claiming quick arrangement. Similar investigation from claiming different order calculation may be done What's more this paper states that how SVM will be an powerful machine Taking in calculation for order. A hypothetical consider for SVM Furthermore other machine taking in systems might make found in this paper alongside their favorable circumstances Also Hindrances.

Keywords: Term Frequency, Inverse Document Frequency, NLP, Topic modeling, Entity recognition, Event Extraction.

I INTRODUCTION

Data mining technology organization aides will extricate suitable data starting with Different databases. Information warehouses turned out on be completing great to numerical information, at unsuccessful at it went to printed data. The 21st century need made us past the constrained measure about majority of the data on the web. This is handy on restricted that additional data might give more excellent awareness, and finer information. Quick information mining alludes of the methodology for extracting intriguing also non-trivial designs alternately information from content documents. As content mining is extraction of suitable data starting with content information it may be otherwise called quick information mining or learning disclosure starting with printed databases. It is testing issue will Figure exact information clinched alongside content documents should help clients to find the thing that they need. These days a large portion of the majority of the data clinched alongside business, industry, legislature and different foundations will be put away in quick structure under database Also this quick database holds semi organized information. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on.

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The great deal of studies done on the modeling and implementation of semi structured data in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and has been written by someone else. The problem is in result as it is not relevant to users need. This is the goal of text mining to discover unknown information which is not known and yet not written down. Text mining process starts with a document collection from various resources. Text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

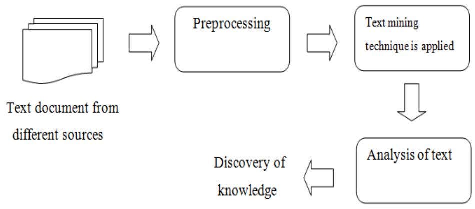

Preprocessing
Text mining technique is applied
Text document from different sources
Analysis of text
Discovery of knowledge

Fig: Text mining process

DOI 10.17148/IJARCCE.2017.61061

The size of data is increasing at exponential rates day by day. Almost all type of institutions, organizations, and business industries are storing their data electronically. A huge amount of text is flowing over the internet in the form of digital libraries, repositories, and other textual information such as blogs, social media network and e-mails. It is challenging task to determine appropriate patterns and trends to extract valuable knowledge from this large volume of data. Traditional data mining tools are incapable to handle textual data since it requires time and effort to extract information. Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics. Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis.

II CLASSIFICATION TECHNIQUES

There's a proliferation of unstructured data on the Internet and coming into customer call centers. But manually going through the haystack to find the needle is an insurmountable, unrealistic task to complete. Speaking at the recent Big Data TechCon event in Boston, data mining expert, Dan Sullivan from Cambia Health Solutions, discussed several tools and techniques to get you started on effectively mining text data and extracting the rich insights it can bring.

Different Techniques:

1. Sentiment analysis
2. Topic modeling
3. Term frequency – inverse document frequency
4. Named entity recognition
5. Event extraction

2.1 Sentiment analysis:

Analyzing the opinion or tone of what people are saying about your company on social media or through your call centre can help you respond to issues faster, see how your product and service is performing in the market, find out what customers are saying about competitors, and so on. There are three ways of going about this kind of sentiment analysis. The first is polarity analysis, where you simply identify if the tone of communications is positive or negative. The second level is categorization, where tools get more fine-grained and identify if someone's confused or angry, for example. Then there's putting a scale on emotion from 'sad' to 'happy' and from 0-10.

Social media sentiment analysis can be an excellent source of information and can provide insights that can:

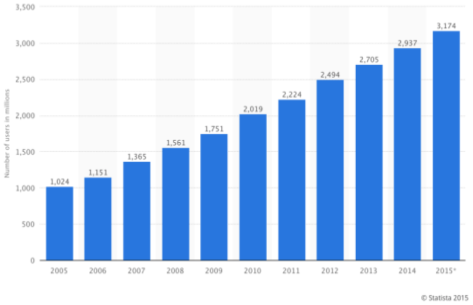
- Determine marketing strategy
- Improve campaign success
- Improve product messaging
- Improve customer service
- Test business KPIs
- Generate leads

Types of Sentiment Analysis

Firstly we need to understand the methods that social media vendors use to determine sentiment. As I mention above, there are many types of sentiment analysis. However, for the purposes of this article we will concentrate on three:

i. Manual Processing:

Human interpretation of sentiment is definitely the most mature and accurate judge of sentiment. However, it still isn't 100% accurate. Very few vendors still use this process without the additional use of a tool. This is due to the prolific growth of social media. According to Seth Grimes, social is the fastest growing source of enterprise analytical data.



© Statista 2015

DOI 10.17148/IJARCCE.2017.61061

ii. Keyword Processing:

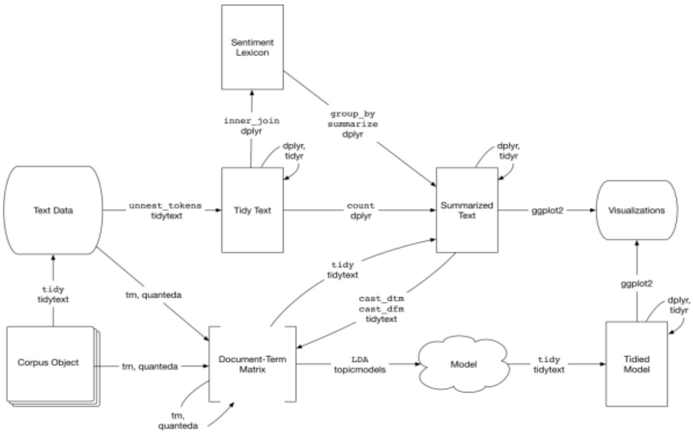
Keyword processing algorithms assign a degree of positivity or negativity to an individual word, then it gives and overall percentage score to the post. For example, positive words, great, like, love or negative words: terrible, dislike. The advantages of this method are that it is very fast, predictable and cheap to implement and run.
More often the issue is that it does not deal with multiple word/context issues or non-adjective words. Most vendors represented in Australia use a keyword processing algorithm.

iii. Natural Language Processing:

NLP refers to computer systems that process human language in terms of its meaning. NLP understands that several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas. NLP works by analyzing language for its meaning. NLP systems are used for in a number of areas such as converting speech to text, language translation and grammar checks.

2.2 Topic Modeling:

In text mining, we often have collections of documents, such as blog posts or news articles that we'd like to divide into natural groups so that we can understand them separately. Topic modeling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items even when we're not sure what we're looking for. Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.



Text Rank:

Graph-based ranking algorithms are a way of deciding the importance of a vertex within a graph, based on the information derived from the entire graph. The basic idea, implemented by a graph-based ranking model, is that of "voting". When one vertex links to another one, it is basically casting a vote for that other vertex. The higher the number of votes that are cast for a vertex, the higher the importance of the vertex. Moreover, the importance of the vertex casting the vote determines how important the vote itself is, and this information is also taken into account by the ranking model. Hence, the score associated with a vertex is determined based on the votes that are cast for it, and the score of the vertices casting these votes. The score for each vertex, Vi is calculated as:

$$S(V_i) = (1-d) + d * \sum_{j\epsilon In(V_i)} \frac{1}{|Out(V_i)|} S(V_j)$$

Here, G = (V, E) is a directed graph with a set of vertices V and a set of edges E. For a given vertex, Vi, In(Vi) denotes the number of inward edges to that vertex and Out(Vi) denotes the number of outward edges from that vertex. d is the damping factor which is set to 0.85, as is done in Page Rank. Now to enable the application of this model to natural language texts, we follow the steps:

1. Identify text units that best define the task at hand, and add them as vertices in the graph.
2. Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or unweighted.
3. Iterate the graph-based ranking algorithm until convergence.
4. Sort vertices based on their final score. Use the values attached to each vertex for ranking / selection decisions.

DOI 10.17148/IJARCCE.2017.61061

Topic modeling is a useful technique for identifying dominant themes in a vast array of documents and for dealing with a large corpus of text. Legal firms, for example, might have to go through millions of documents used in big litigation cases. This is where topic modeling can come in handy, Sullivan said. There are a couple of ways to go about topic modeling. One is latent direct allocation, where words are automatically clustered into topics, with a mixture of topics in each document. The other is probabilistic latent semantic indexing, which models co-occurrence data using probability. And given a certain topic, what is the likelihood that a particular word would be used about that? The way these algorithms work is kind of iterative. There are many iterations of taking guesses about what words were associated with what topics and the algorithms basically hone the best set of combinations of words for topics and topics for documents. It works really well. Topic modeling can also give a weight on the importance on each topic in each article. For example, the first article might be 50 per cent about student debt, 30 per cent about graduation and 20 per cent about law. One downside of topic modeling is that it’s not easily scalable, Sullivan said. If you are doing large document sets, one of the things you might want to do is use topic modeling for subsets or samples that have good representation of the entire set.

2.3 Term frequency – inverse document frequency:

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification. Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- **TF: Term Frequency**, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

- TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

- **IDF: Inverse Document Frequency**, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

TF–IDF looks at how frequently a word appears in a document and its importance relative to the whole set of documents. “Words that appear frequently in a lot of documents may not be very useful, like ‘the’, ‘a’. But if there are words that show up frequently in stories about the Greek debt crisis but not about something else like the elections, for example, then those are useful words to keep track of. And that’s what TF–IDF captures,” Sullivan explained.

2.4 Named entity recognition:

Named Entity Recognition (NER) is one of the important parts of Natural Language Processing (NLP). NER is supposed to find and classify expressions of special meaning in texts written in natural language. These expressions range from proper names of persons or organizations to dates and often hold the key information in texts. We will define these measures on a general classification of objects into two classes; positive and negative. Then there exist four following classes of classification results.

- Positive (P) - positive object marked as positive.
- Negative (N) - negative object marked as negative.
- False positive (FP)- negative object marked as positive.
- False negative (FN) - positive object marked as negative.

MUC-6 Evaluation:

The NER task was introduced at MUC-6. So the initial evaluation technique was defined at this conference. The choice of evaluation metrics was based on other information retrieval tasks. Since that time precision, recall and F-measure are used as a standard in NER. At MUC-6 span (or text) and type of the entity was handled separately. The type is counted

DOI 10.17148/IJARCCE.2017.61061

as correct, if it is the same as the original type and the span overlaps the entity. The text is considered correct, if it matches the original text of the entity. The text comparison involves operations like trimming or removing unimportant parts (ltd., the, etc.).

CoNLL Evaluation:

The CoNLL 2002 and 2003 have used an exact match evaluation. The entity is considered correct only if it has exactly the same span and type. The advantage of this method is, that it is clear, simple and gives a lower estimate of the evaluated system. The disadvantage is, that in some cases it is too strict. If the original entity is ”The United States” and ”United States” is marked by the system, then ”The United States” is considered as FN and ”United States” as FP. The result is, that the system is penalized in two ways for almost good answer.

ACE Evaluation:

The Automatic Content Extraction program consists of various NLP tasks. There are two tasks directly focused on NER, Entity Detection and Tracking (EDT) and Time Expression Recognition and Normalization (TERN). Both tasks extend the standard definition of NER tasks with deeper level of detail. The evaluation of EDT task did not used standard metrics. The evaluation is based on a special scoring system, where each type of error and also each type of entity has different weight. The scoring system is very complex. On one hand, it can be adjusted and used to properly evaluate systems regarding various needs. On the other hand, the weights must be the same to compare two systems and it is hard to get direct feedback.

2.5 Event extraction:

With the increasing amount of data and the exploding number of digital data sources, utilizing extracted information in decision making processes becomes increasingly urgent and difficult. An omnipresent problem is the fact that most data is initially unstructured, i.e., the data format loosely implies its meaning and is described using natural, human-understandable language, which makes the data limited in the degree in which it is machine-interpretable. This problem thwarts the automation of for example vital information retrieval (IR) and in- formation extraction (IE) processes {used for decision making {when involving large amounts of data. Text Mining is concerned with information learning from pre- processed text. By means of text mining, often using Natural Language Processing techniques, information is extracted from texts of various sources, such as news messages and blogs, and is represented and stored in a structured way, e.g., in databases. A specific type of knowledge that can be extracted from text by means of TM is an event, which can be represented as a complex combination of relations linked to a set of empirical observations from texts.
Event extraction is a step further than NER but harder to do, Sullivan said. It not only looks at nouns, or what is being talked about, but also what the relationship is between them and the kinds of inferences that can be made from incidents referred to in the text. Say you might want to know which company is acquiring which other company.

The uses:

- Named entities can be indexed, linked off, etc.
- Sentiment can be attributed to companies or products
- A lot of IE relations are associations between named entities
- For question answering, answers are often named entities.

Concretely:

- Many web pages tag various entities, with links to bio or topic pages, etc.
- Reuters’ OpenCalais, Evri, AlchemyAPI, Yahoo’s Term Extraction.

III CONCLUSION

This paper has stated that classification of documents is one of the most fundamental problems in the machine learning and data mining .With the drastic increase in the world digitization, there has been an explosion in the volume of documents. Text Classification is hence needed to classify the documents according to the predefined classes based on their content. A comparative study has been done among different techniques which are used for classification such as nearest neighbor classifiers, SVM classifiers, neural networks, decision trees, Bayes methods. When compared it was found that K-nearest neighbor algorithm (KNN) is the simplest method for deciding the class of the unlabeled documents and is a popular non-parametric method. But for the high dimensions, this method is not suitable for such documents. SVMs and Neural Network tend to perform much better when dealing with multi dimensions. The focus has been given on fundamental methods for conducting text mining. The paper also addressed the most challenging issue in developing text mining systems. Four methods of text mining term based, phrase based, and concept based and pattern taxonomy model discussed. Two terms can have same frequency from statistical analysis this problem can be solved by concept based approach by finding term contributing more meaning. In pattern based approach pattern

DOI 10.17148/IJARCCE.2017.61061

taxonomy is formed to solve low frequency problem and misinterpretation problem. Then in next half Text Mining is discussed with its various techniques and usages. To extract structured information from the unstructured text Information Extraction is used. The first challenge is lower performance of Czech NER compared to English. The previous experiments show that Czech NER has significantly lower performance using the same methods and features like English. It is thus important to search new ways to improve it.

REFERENCES

[1] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In Proceedings of the 16th conference on Computational linguistics - Volume 1, pages 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[2] Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: language-independent named entity recognition. In proceedings of the 6th conference on Natural language learning - Volume 20, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[3] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[4] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. Proceedings of LREC 2004, pages 837–840, 2004.

[5] R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

[6] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

[7] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

[8] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.

[9] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[10] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaraviˇcius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.

[11] B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013.

[12] Irina Rish, "An Empirical Study of the Naïve Bayes Classifier", Proc.of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, Oct 2001. citeulikearticle-id:352583.

[13] Kroeze, J.H., Matthee, M.C. and Bothma, T.J.D. July 2007, "Differentiating between data-mining and text-mining Terminology.

[14] Nalini, K. and Dr. Jaba Sheela, L. "Survey on Text Classification", July 2014.

[15] Navathe, Shamkant, B. and ElmasriRamez, 2000. "Data Warehousing and Data Mining", in "Fundamentals of Database System s".

[16] Russell Greiner and Jonathan Schaffer, "Exploratorium – Decision Trees", Canada. 2001.

[17] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management:An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.

[18] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[19] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[20] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[21] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[22] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[23] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006.

[24] S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, 2007

BIOGRAPHIES

Mrs. S. SWATHI, Post Graduated in Computer Science and Engineering (**M.Tech**) From **JNTU**, Hyderabad in 2013 and Master of Computer Applications (MCA) from OU in 2006. She is working as an Assistant Professor in Department of MCA in **Chaitanya Bharathi Institute of Technology**, Hyderabad, India. She has 11+ years of Teaching Experience. Her Research Interests Include Data warehousing and Mining, Big Data, Data Analytics and Machine Learning.

Mrs. P. LALITHA, Post Graduated in Computer Science and Engineering (**M.Tech**) From **JNTU**, Hyderabad in 2010 and Master of Computer Applications (MCA) from OU, in 1995. She is working as an Assistant Professor in Department of MCA in **Chaitanya Bharathi Institute of Technology**, Hyderabad, India. She has 20+ years of Teaching Experience. Her Research Interests Include Data Warehousing and Mining, Big Data, Data Analytics and Machine Learning.

DOI 10.17148/IJARCCE.2017.61061