An Application of Data Analytics for Social Media Platforms and e-Governance: An Overview

Dhanashree Munot, Dhiraj Wakharde, Vrushali Kamble, Charudatta Potdar, Sushma Vispute, Rajeswari Kannan

Pimpri Chinchwad College of Engineering, Pune, India

Corresponding author: Dhanashree Munot, Email: dhanashreemunot106@gmail.com

Social media is a great platform that contains a pool of information combined with people. This can be used to analyze data for better results for an organization. The objective of the paper is to study data analytics and how this concept can be used in social media text analysis. Natural language processing is used for text analysis, information extraction, etc. This survey paper mainly focuses on understanding the need for data analytics in text analysis through various domains like governance, politics, and rural development. Classification and clustering techniques derived from machine learning that helps to extract important information through various algorithms to help companies, organizations, governing bodies understand their audience properly and work according to their needs. The paper consists of a detailed analysis of clustering and classification algorithms over a wide variety of domains and compares the results of the performance metrics. According to the research done by the authors, it is found that Fine-tuning BERT classifier gives the highest accuracy among other classification algorithms which includes Naïve Bayes Classifier, Decision Tree, and Support Vector Machine (SVM). The study shows that clustering has been used for finding sets of similar words, sentences, word sense, etc. These concepts can be used for solving problems related to e-governance as these authorities can deploy such methods to understand their people and work according to their needs.

Keywords: Machine Learning, Data Analytics, Clustering, Classification, Performance Metrics.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 261–270. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-28

1 Introduction

There is a huge amount of data on Social Media platforms that can be utilized to predict what the citizens want from the government. This data has an enormous potential to give useful insights that can benefit both the citizens and the smooth functioning of the governing bodies. The infusion of technology to read this data, understand the data patterns, draw conclusions, and predict future trends is required. We are trying to exploit the power of Natural Language Processing to understand such data. We are surveying different algorithms and comparing their accuracy for such problem statements.

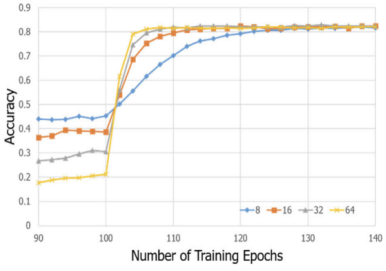
Some common terminologies that are required for this survey are listed below:

Data Analytics: Data Analytics method of scrutinizing unanalyzed data and utilizing it to draw some important conclusions. It happens through a series of steps including finding data requirements, collecting the data, organizing data, cleaning the data, and finally analyzing data for results and conclusions. Types of data analytics include - Descriptive, Predictive, Prescriptive, and Diagnostic.

NLP: Natural language processing is a sub-domain of Artificial Intelligence where analysis techniques of language interaction between computers and humans present it in an understandable form. Two parts of NLP: data pre-processing and algorithm generation. Data pre-processing deals with the preparation of data and cleaning data that will help machines to analyze the data. After pre-processing, the development of algorithms is done which are mainly rule-based systems and machine learning-based systems.

2 Literature Survey

Xuan et al. [1] have used details of labels in a pair wise way, and with the help of this, they have proposed a semantic text hashing method that is supervised in nature. For the calculation of similarity in text pairs in a pair wise fashion, numerous ways which are dependent upon the auto-encoder model are used. Now as the calculation of similarity does not require much time, the full studying process is boosted up and better organized than those in the existing methodologies. It is seen that the results after experimentation that made use of open datasets showed that the proposed process can effectively utilize pair wise information of labels and give more favorable outcomes than more antiquated state-of-the-art approaches of hashing.



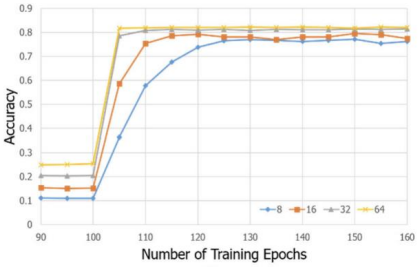


Fig. 1. Accuracy for each training epoch, Source[1], **Fig. 2.** Accuracy for each training epoch, Source [1]

The above two figures Fig.1 and Fig.2 representations portray the accuracy for every training epoch with unlike group dimensions when utilizing the 20 newsgroups dataset.

(a) VPSH-Bernoulli.
(b) VPSH-Gumbel.

Jayo and Almeida [2] have proposed a new way of Political discourse analysis. They have used political manifestos and Twitter both as their experimental subjects. Combining these two datasets has yielded better results.

Stier et al. [3] have used machine learning models to monitor politician's social media accounts. Based on what politicians tweet and post on Facebook the machine learning model identifies whether the politicians are addressing the problems of the masses. This is used for election campaigning.

Ripollés [4] surveyed how social media is changing the manufacture, dispensation, and usage of political information. The author has also stated the areas that are open for research in this area.

Mendhe et al. [5] have talked about how the past few years Twitter has been quite a successful and well-known social media platform. Many researchers and students need twitter data to build machine learning models. But the Twitter API has its limitations. Getting access to complete data can be very expensive for students. That's why the authors have come up with an AI platform where, with a click of a button, anyone can have access to large amounts of data and can build their AI applications.

Sánchez et al. [6] have spoken about how the fast evolution of web technologies and social media has authorized the exchange of thoughts between people all over the world and socio-economic cultures. The abhorrent and chauvinist content concerning women is being progressively spread on social networks. They have focused on recognizing how chauvinist attributes, credence, and viewpoints are shown in discussions on Twitter. They proposed to recognize chauvinism using machine learning techniques. The sequence of machine learning algorithms was used to separate into sexist or nonsexist tweets. It was observed that chauvinism and hatred towards women were detected more easily than subtle chauvinism. They discussed the performance of automatic methods of classification that were concerned with all the types of chauvinism. The investigational outcomes showed that BERT extraordinarily performs better than other algorithms, getting an accuracy of 74% in the detection of chauvinist voicing. It was shown that a classification system that was made to train on their dataset was better generalized than the same system that was trained on a dataset of misogynistic expressions.

Bose et al. [7] have talked about the devastating effects of COVID-19 on the lives of people. Their work had an application of Natural Language Processing-based approaches to infer significant keywords which gave their part to the social, economic, clinical, medical understanding of this devastating pandemic. They also managed to figure out countries, headings, and research artifacts that portrayed that the science body still reacted to the short-term dangers such as fitness dangers, therapy planning, policies, etc. Thus, surveying written data of published artifacts in these fields could underline research gain grounds in virus-borne diseases in general and COVID-19 in specific. This helped to inform future research, clinical trials, treatment methodologies and course, implications, and administrative decision-making. The authors used the variation coefficient concept to search essential keywords both semantically and analytically and use this to find trends in the entire research of COVID-19.

The Fig. 3 representation from the paper mentioned the occurrence of 25 most introduced countries in COVID-19 science.

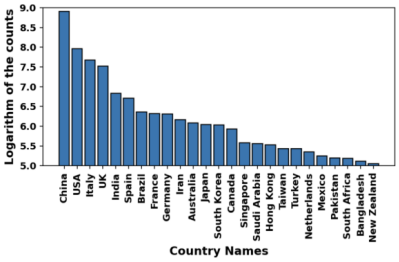


Fig. 3. Occurrence of 25 most introduced countries in COVID science. Source [7]

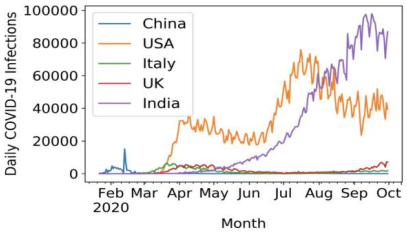


Fig. 4. Everyday COVID-19 contaminations in the 5 countries introduced maximum in COVID-19 sciences Source [7]

The above Fig. 4 representation from the paper mentioned the everyday COVID-19 sciences. It showed an analysis of nine months in the year 2020 as to how many patients were detected in countries like Italy, the UK, China, the USA, and India. It shows a clear representation of the statistics as to where the most infected people lived.

Kane [8] proposed a way to attract augmentations with modernized tools of AI in governance. The authors gave a methodology that could assist electorates around certain vicinity to build up notable uprightness within their election rotations. This will in turn result in an increasing level of trust in the institutions of society along with fair election conditions.

Wan et al. [9] have discussed the current expansion swing which is leading to intensified requests for intelligent technologies to lead numerous candidatures. Here they developed a mechanized NLP-based framework which led to empowering and complementing congestion announcement solutions. This consisted of social media text mining, data extraction, and giving alerts to drivers. They have employed the fine-tuning Bidirectional Encoder presentation from Transformers classification model to filler and for classification of information. A Question-Answer model was then followed which extracted required data that helped in characterizing accidents that were reported. The data fields contained the place, time, and the incidence of nature. This data was then converted to alerts which were unified into personal navigation assistants. In the end, a comparison between the recent reports of all incidents from official authorities and also social media was done so that a clearer picture is observed and also suggested some investigation pathways.

Akhter et al. [10] have shown the Machine Learning models were employed in the processing of text. According to them, this is the first examination of Urdu TDC utilizing a deep learning model. They have designed a hugely adaptable and assorted dataset that comprises thousands of documents categorized into six types. They used SMFCNN (Single-layer Multisize Filters Convolutional Neural Network) for classifying and collating its presentation with different Machine Learning models. Additionally, they

inspected the outcome of pre-processing techniques on SMFCNN performance. They designed such a dataset that will be publicly accessible in various formats for further investigation.

Balcı and Salah [11] put some light on Machine Learning techniques and Artificial Intelligence techniques which are used for creating reasonable attributes for bilateral elements of the game. They also analyzed the players' attributes which provided a better gaming habitat. They have proposed a novel skeleton for player classification that is automated in a social gaming program. They demonstrated characteristics that described both parties of the objection and also game features that are interacting. They used an approach that was classification, which was based on inclination boosting machines. They showed a method to mechanically identify authentic player objections for oral abusive content and insulting attributes in an online communal game program. They also showed that the extension characteristics set performs great, the usage of only player profiles that have a suspicion.

Castillo and Flores [12] demonstrated a survey in Music Data Extraction consisting of a wide range of themes consisting of genre categorization, locating, counsel and visualization. They referred to comprehension from music and involved its investigation and study. They showed wonderful results when they combined the Machine Learning approach, helping in emulating human abilities which in turn helped the final user. The usage of a web application demonstrated recovering musical songs from a platform like YouTube which assorted all of them in different musical categories. The usage of classifiers from distinct ML guidelines like the RNN, Feed-forward, and many more such prototypes. All these prototypes were then trained in numerous-label categorization plans. The theory and science behind these proposed problems of classifiers were thoroughly explained. They used three different musical notes as their case studies and portrayed their application in real practice. Analysis of online classifications was done to further talk about the model efficiency results.

Kausar et al. [13] stated that the tweets of people played a vital role in sentiment analysis of the countries affected to its maximum extent. The collected data from the maximum infected countries and also one of the Gulf region countries went through experimental processes. A tweet analysis was conducted in the English Language. On this survey, an emotion-based analysis was executed. The sole purpose behind this experimentation was to understand the circumstances faced by the people who were infected. The usage of text mining algorithms, preprocessing of the gathered data of the tweets took place. The purpose was to know the emotions of those infected by the COVID-19 disaster.

Kim and Lim [14] stated that scanning customer complaints were bestowed by a data-driven method. This was used for systematic management of the service quality. The worth of the customer opinion was done by a poll of "Voice of the customer". The merged methodologies of emotions and analysis of the stats process were proposed in that poll. The customer review information for the stats process control survey was used. The implementation from supplier perception to customer-focused management of service was made from the above poll. The surveys were then integrated to make it look simpler for understanding customer dissatisfaction points. These were then accepted concerning the cost spent and the time consumed. The proposed methodology was validated and applied. They used a mobile service, which helped them achieve guidelines and prototypes for the sake of execution and alteration if any. This process should be responsive and preventive quality management is expected.

The overall System process is explained in Fig. 5

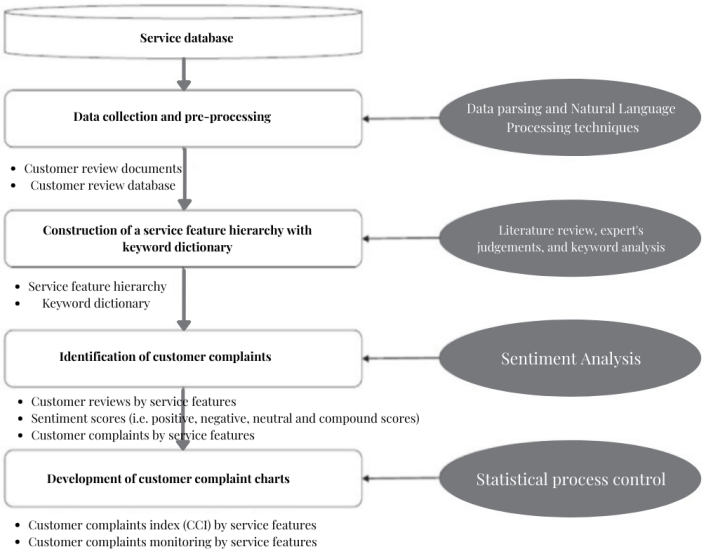


Fig. 5. Customer Complaint Analysis Source[14]

This system consists of the following steps:

1. Collection of data and data pre-processing using data parsing and NLP techniques.
2. Using the collected data, build a service with a keyword dictionary.
3. Complaints are recognized that are submitted by customers using sentiment analysis.
4. Complaint charts are developed using SPC.

Mills and Bourbakis [15] performed research that portrayed the analysis and survey of the performance, functional components of the algorithms, and the graph-based methodologies for Natural Language Processing. They showed the potential for developed and matured products which were obtained by natural language understanding. They concluded by presenting the results of the surveys which included text entailment, similarity measure, review, redundancy decreasing, similarity measure, semantic relativeness, labeling, disambiguation, and word sense induction. They offered the estimated scores for performance, the usability of algorithms, accuracy, and scalability from each methodology. Bar graphs and tables were used to prove maturity level and functional component abstraction. Various clustering methodologies comparison was done by the authors as shown in Table 1.

Table 1. Clustering Methodologies, Source [15]

Methodology	Performance
Similar Words Clustering and Concepts Hierarchies	Accuracy: Noun (90.0%), Verb (77.6%), Adjective (92.2%)
Classifying Context Dependencies Of the Parsed Annotated Text	F-score 41%–52% in detecting and identifying events
Metrics based on Graphs to filter and delete error terms	89% accuracy
Map-Reduced Algorithm for Label Propagation	Noun (58%), Verb (83%), Adjective (73%) F-measured

Bertot and Jaeger [16] have talked about the techniques of research and evaluation of social media which initiates and encourages the usage, continually tool improvisation, updating the policies, controlling their usage, and processes of governance development that incorporate the participation of social media in different ways.

Hu et al. [17] have talked about a study that shapes an E-Governance Services point of influence factor in the form of a model of public engagements. The technology acceptance theory, motivation theory, and trust theory led to exploring and finding the impact reasons and pathways for resolving problems. Comprehension validation was conducted through impact path surveying which used Structural Equation Modeling (SEM) methodologies. The analysis showed that a particular way the public accepted and adopted the co-creation value behavior for E-Governance Services. This research is surely helpful for more effective incentives mechanisms and for IT managers to enhance E-Governance Services value creation.

Ding et al. [18] have talked about how the triumph of mobile phone microblogging services is directly censorious to people's voices and their opinions. It made use of gratification theory and the impetus organism response framework. In their developed research model, the investigation of the effect of discerning integration and the nature of the gratification was implemented. This had an impact on the mobile government microblogging services. Their survey of neural network research portrayed the effect of nature on information values and the hedonic value. These hedonic values were shown more powerful than that of the recognized integration. The theoretical and managerial suggestions for mobile phone government microblogging services were their conclusion points.

3 Algorithmic Survey

3.1 Binary classification

Wan et al. [9] proposed a model for classifying social media data in two traffic-related collections. For building this, the model used is a fine-tuned BERT classifier model. It is a pre-trained model but due to fine-tuning of over 340 million parameters in the original BERT large model, this model yields an accuracy of 98.9% as shown in Fig. 6. The confusion matrix consists of true positive (true_pos) , false positive(false_pos), false negative(false_neg) and true negative(true_neg) rates. Matthews's correlation coefficient (MCC) is calculated to evaluate the accuracy, precision, and recall metrics.

To achieve the accuracy of 99.6% as shown in Fig.6, first a group of 106437 tweets is divided into a training set of 80% and a testing set of 20%. The rate of learning for optimizers is set to 2×10^(−5) and the dropout rate to 0.1.

The optimization method used is Gaussian Error Linear Unit (GELU) which has five hidden layers.

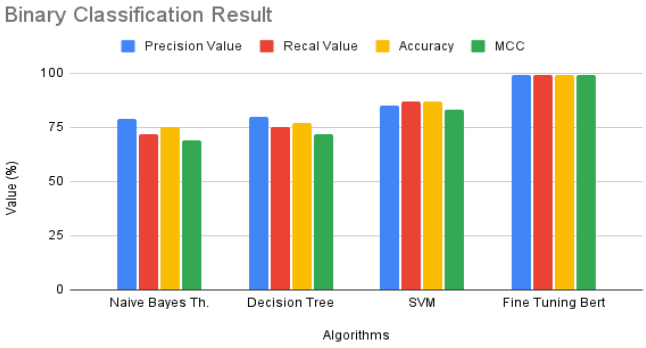


Fig. 6. Binary classification algorithms and results

3.2. Multiclass Classification

Wan et al. [9] used a multiclass classification algorithm to classify traffic-related social media input data into multiple categories. When the output layer of the BERT model is changed, it becomes a multiclass classification algorithm.

1. The data is divided into six categories.
2. The learning rate is set to 2 × 10^−5.
3. The size of the batch is 24.
4. Gaussian Error Linear Unit (GELU) activation function is used.
5. The Classifier categorizes data into 6 different classes namely Incident, Construction, Public transportation, Delay, Closure, and Unrelated.

The accuracy reached by this algorithm is 99.37% as shown in Fig. 7. To achieve this accuracy the model is trained on a training set of 29418 and tested on a test set of 7357 tweets. The batch size is set to 24 with a dropout rate of 0.1. The function drops rapidly in the initial 200 epochs and slowly becomes steady.

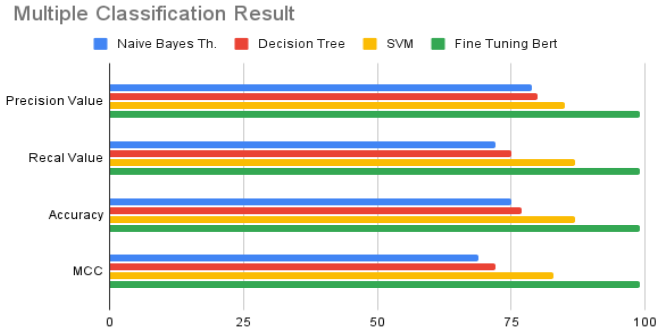


Fig. 7. Multiple Classification algorithms and results

4 Results

According to the survey which we have done regarding the use of NLP to study social media posts, customer sentiments, classification can be summarized in the following Table 2 and the percentage of each type of paper is shown in Fig. 8.

Table 2. Survey Paper types and common findings

Research Paper Type	No. of paper referred	Common Findings
Political	5	Machine learning models have been used to check the social media accounts of politicians or find their activities through social media and relate them to people and their governance.
Twitter	4	Twitter has been for various sentiment analysis work to check the impact which is created on people through the data posted on Twitter.
E-Governance	4	The relationship between rural development and technological advancement has been studied and all measures are required to be taken using AI and social media data for the development of such areas.
Natural Language Processing	5	The use of AI and Machine Learning has been done for text data processing majorly.
Other	4	Various classification and clustering techniques have been used. For example, Automatic players' complaint classification, Automatic music genre classification, etc.

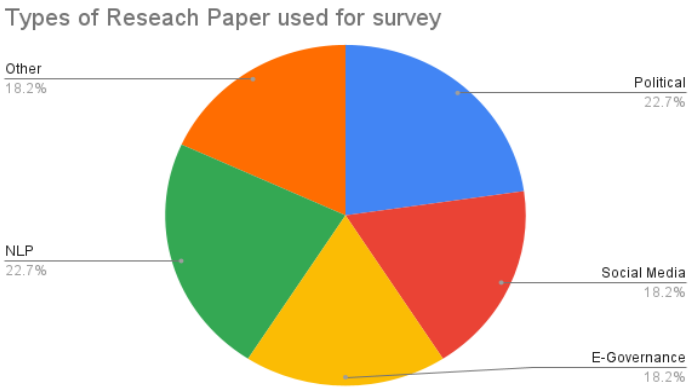


Fig. 8. Types of paper and their weightage

5 Conclusion

In this research paper, a study on how to use NLP to analyze social media posts and know what the thoughts are, and generate insights from them has been done. Data Analysis techniques for better decision-making about emerging trends. Sentiment analysis is an important metric as it indicates people's favorability about certain trends, events, etc. The studies showed that Multi-Classification accuracy for various algorithms is between 75% to 99% and for binary classification accuracy is 94% to 99% for algorithms like Naive Bayes, SVM, Fine-tuning BERT.

The studies showed wonderful results when it was combined with the Machine Learning approach, helping in emulating human abilities which in turn helped the final user. The main objective behind the analysis paper was to discuss the impact of AI and social media in rural areas for development purposes. The studies put some light on the relationship of Artificial Intelligence to improve rural areas by the implementation of various technologies related to social media.

References

[1] Xuan, R. et a. (2021). Deep Semantic Hashing Using Pairwise Labels. *IEEE Access*, 9: 91934-91949.
[2] Jayo, A. B. and Almeida, A. (2021). Improving Political Discourse Analysis on Twitter with Context Analysis. *IEEE Access*, 9: 104846-104863.
[3] Stier, S. et al. (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, 35: 50–74.
[4] Ripollés, A. C. (2018). Research on political information and social media: Key points and challenges for the future. *El Profesional de la Información*, 27(5): 974-974.
[5] Mendhe, C. H. et al. (2021). A Scalable Platform to Collect, Store, Visualize and Analyze Big Data in Real Time. *IEEE Transactions on Computational Social Systems*, 8(1): 260-269
[6] Sánchez, F. R. et al. (2020). Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access*, 8: 219563-219576.
[7] Bose, P. et al. (2021). A Comparative NLP-Based Study on the Current Trends and Future Directions in COVID-19 Research. *IEEE Access*, 9: 78341-78355.
[8] Kane, T. B. (2019). Artificial Intelligence in Politics. *IEEE Technology and Society Magazine*.
[9] Wan, X. et al. (2020). Empowering Real-Time Traffic Reporting Systems with NLP-Processed Social Media Data. In *IEEE Technology and Engineering Management Conference*.
[10] Akhter, M. P. et al. (2020). Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network. In *IEEE Access*, 8: 42689-42707.
[11] Balcı, K. and Salah, A. A. (2017). Automatic Classification of Player Complaints in Social Games. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(1): 103-108.
[12] Castillo, J. R. and Flores, M. J. (2021). Web-Based Music Genre Classification for Timeline Song Visualization and Analysis. *IEEE Access*, 9: 18801-18816.
[13] Kausar, M. A. (2021). Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak. *International Journal of Advanced Computer Science and Applications*, 12(2): 415-422.
[14] Kim, J. and Lim, C. (2021). Customer complaints monitoring with customer review data analytics: An integrated method of sentiment and statistical process control analyses. *Adv. Engg. Informatics*, 49: 101304.
[15] Mills, M. T. and Bourbakis, N. G. (2014). Graph-Based Methods for Natural Language Processing and Understanding—A Survey and Analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(1): 59-71.
[16] Bertot, J. C. and Jaeger, P. T. (2014). Social Media Technology and Government Transparency. *IEEE Computer Society*.
[17] Hu, G. et al. (2019). The Influence of Public Engaging Intention on Value Co-Creation of E-Government Services. *IEEE Computer Society*.
[18] Ding, Y. et al. (2019). Explaining and Predicting Mobile Government Microblogging Services Participation Behaviors: A SEM-Neural Network Method. *IEEE Computer Society*.