A Multi-Strategy Approach for Information Extraction of Financial Report Documents

Siti Mariyah
Department of Statistical Computation,
Sekolah Tinggi Ilmu Statistik, Jakarta
Email: sitimariyah@stis.ac.id

Dwi Hendratmo Widyantoro
School of Electrical Engineering and Informatics,
Institut Teknologi Bandung, Bandung
Email: dwi@stei.itb.ac.id

***Abstract*—Information extraction studies have been conducted to improve the efficiency ansd accuracy of information retrieval. We developed information extraction techniques to extract name of company, period of document, currency, revenue, and number of employee information from financial report documents automatically. Different with other works, we applied a multi-strategy approach for developing extraction techniques. We separated information based on its similar characteristics before designing extraction techniques. We assumed that the difference of characteristics owned by each information induces the difference of strategy applied. First strategy is constructing extraction techniques using rule-based extraction method on information, which has good regularity on orthographic and layout features such as name of company, period of document and currency. Second strategy is applying machine learning-based extraction method on information, which has rich contextual and list look-up features such as revenue and number of employee. On the first strategy, rule patterns are defined by combining orthographic, layout, and limited contextual features. Defined rule patterns succeed to extract information and gain precision, recall, and F1-measure more than 0.98. On the second strategy, we conducted extraction task as classification task. First, we built classification models using Naïve Bayes and Support Vector Machines algorithms. Then, we extracted the most informative features to train the classification models. The best classification model is used for extraction task. Contextual and list look-up features play important role in improving extraction performance. Second strategy succeed to extract revenue and number of employee information and gains precision, recall, and F-1 measure more than 0.93.**

Keywords*— information extraction, contextual features, orthographic features, list lookup features, Naïve Bayes, Support Vector Machines

I. INTRODUCTION

The advent of internet technology has influenced many aspects of life includes business aspect. The large amounts of digital data have been created due to use of technology internet, e.g. financial data. An increase of digital data has triggered the birth of many search engines. Search engines help users to find information and filter unrelated documents in the internet resources. However, some users' specific needs are not met by searching for data using keyword retrieval technique [1]. Users must read each delivered documents to meet their specific data. For example, a routine activity conducted by one of government institutional. Before they release Tendency Business Index [2], they are always looking for financial data on internet to get the supporting data about the business condition. They read all documents, which are provided by search engine and filter unrelated financial data manually. This routine task is time consuming. An innovation is needed to handle the task in order to be effective, efficient, accurate and automatic.

Information extraction is a solution to improve the effectiveness and accuracy of information retrieval. Information extraction meets specific users' needs by providing specific and related information. Information extraction in the financial domain is expected to handle the task and to provide the accurate financial information. In general, information extraction has two methods, rule-based and machine learning-based extraction method. Rule-based extraction method requires human experts to define accurate rules or program code for performing extraction task. Sarawagi [3] said, "Person needs to be a domain expert and a programmer, and possess descent linguistic understanding to be able to develop robust extraction rules". While machine learning-based method requires labeled unstructured examples (text) to train machine learning models of extraction. Each method is applied depends on the characteristics of each information to be extracted.

Some existing studies of information extraction applied single extraction method on all target information (information to be extracted). They tried to find the best rules or the best machine learning models for extraction task. They assigned a single method for extracting information before identifying and analyzing the characteristics owned by each target information [4][5][6][7]. This research developed extraction techniques to extract financial information from financial report documents, which consist of name of company, period of document, currency, revenue and number of employee. Different with others, this research built extraction techniques by applying a multi-strategy approach of information extraction. We identified and analyzed characteristics before assigning an appropriate strategy to extract information.

Target information have its own characteristics. Name of company, period of document and currency have a good regularity while revenue and the number of employee do not.

One approach cannot be applied for all target information. Therefore, we designed a multi-strategy approach that tailored to the individual characteristics then developed the appropriate extraction techniques. The rest of paper is organized as follows: section 2 describes related works; section 3 explains information extraction techniques; section 4 discusses experiments and results; and section 5 presents conclusion and future works.

II. Related Works

Some works of information extraction in the financial domain have been conducted in [1][4][5][6][7]. J. L. Seng and J. T. Lai [1] developed extraction techniques on financial data from Chinese-based HTML pages and PDF files. Financial data are delivered on financial statements, notes to financial statements and financial news for business valuation. They used intelligent word segmentation, lexical analysis module, and lexical extraction module to develop automatically extraction technique. This research remains some problems such as extraction design, extraction methods for different data sources, and integral business valuation database. Multiple binary classifier and rule-based extraction methods are used for extracting dividend event in Europe press release corpus [4]. Rules are not used to extract information directly but to fix erroneous resulted by multiple binary classifiers. This research gained precision 96% and recall 79%.

P. Andre and S. Ratte [5] built acronym extraction technique using machine learning-based extraction method on France business documents. Business documents have different characteristics with biology and technical documents. Therefore, acronym extraction on biology and technical documents cannot be adopted by business documents. Implicit presentation of acronyms, the proximity of the acronym's form, and presentation devices to be challenges on extraction process. Extraction method needs deep syntactic analysis for finding troublesome of business documents. P. Andre and S. Ratte introduced similarity features where its values are from comparing of candidate characteristics with average length calculated from acronym repository. This technique gains *precision* 89.1% and recall 90.9%.

M. Sheikh and S. Conlon [6] introduced rule-based extraction method to extract financial information consists of financial factor, previous financial factor, current volume, previous volume, change type, and change volume for investment decision making. These methods include symbolic learning method trained using Greedy Search and similarity model trained using Tabu Search. Developed system only consider presence or absence of feature constraint for generalizing two different extraction model. First extraction model developed using Greedy Search method and second extraction model developed using Tabu Search method. Research by Hui Han, C. Lee Giles, E. Manavoglu, and Hongyuan Zha [7] explored contextual features to extract documents metadata from digital library such as Citeseer and EbizSearch. Contextual features are number of punctuation, number of words after punctuation, number of words before punctuation, and number of words between punctuation.

III. Information Extraction Techniques

In information extraction, characteristics of information are represented in features. We have explored several types of features that are most likely owned by any target information. We selected the most informative features. The most informative features consist of orthographic, contextual, layout, and list look-up features.

Orthographic features – we investigated that name of company, period of documents, and currency have clear orthographic features such as use of capital letter, punctuation, and symbols. Name of company token consists of three part, token_begin is combination of 'PT', token_in consists of more than two words, and token_end is combination of 'Tbk'. Period of document has format "dd(date) MM(month) yyyy(year)". Currency has values are 'rupiah', 'ribuan rupiah', 'jutaan rupiah', 'US$', 'ribuan US$', 'US$ 000', dan 'Rp '000''. Name of company, period of document, and currency are written in title of tables or sometimes for currency token is written in first row of tables. Revenue and number of employee token do not have special orthographic features.

Contextual features- we found that name of company and period of document token do not have clear contextual features. Currency, revenue, and number of employee token have contextual features such as unigram before, unigram after, bigrams before, bigrams after, previous line, next line, and type of unigram after and before. Currency token is preceded by phrases "expressed in" or "presented in". Revenue token is preceded by 20 terms such as "consolidated profit before income tax", "operating revenues", "income tax benefit/expense", etc. and followed by numeric token. Number of employee token is preceded by phrases "number of employees of", "the company has", etc. and followed by phrases "people and", "employee and", and "permanent employees".

Layout feature – name of company token is written on first row of title of table. Period of document and currency are written on second or third of table title. While revenue token is written in different tables such as table "Notes to the Consolidated Financial Statements", "Consolidated Statements of Comprehensive Income", etc. From 939 documents, we found that there are three formats of table. Formats consist of one period table format, two period table format, and four period table format. Table formats influence contextual features of revenue token. While number of employee token is written in paragraph.

List Look-up features- we identified all words or phrases that can characterize revenue and number of employee tokens.

We identified four list look-up features for revenue token and five list look-up features for number of employee token.

Table 1. List Look-up Features for Revenue Token

No	Type	Content
1	Unigrams as revenue identifier	[consolidated, interim, income, tax, before, comprehensive]
2	Bigrams as revenue identifier	[income tax, consolidated interim, before tax, consolidated comprehensive, group, operating revenues, comprehensive income, cooperate income, comprehensive loss, net loss, consolidates loss, (benefit) tax, operating profit]
3	Unigrams as non-revenue identifier	[total, segment, entity, parent, deferred, premium, administration, sales, donation, fiscal, work, customers, equity, reserved, asset, service, operating, rest/etc.]
4	Bigrams as non-revenue identifier	[parent entity, current tax, deferred tax, premium income, segment results, sale expenses, sale, amounts of revenue, post-employment, unappropriated, amounts of equity, amounts of assets, payments of services, fiscal losses, financial income, amounts of sales, operating benefits]

Table 2. List Look-up Features for Number of Employee Token

No	Type	Contents
1	Previous phrases of number of employee	[fixed, as many as, company has, company had, employees]
2	Next phrases of number of employee	[people and, employees and, people, and employees respectively, by employee, and employees, permanent employees]
3	Bigrams as non-number of employee identifier	[to employees, welfare of employees, for employees, salary of employees, all employees, employee services, bonuses of employees, benefit of employees, cost of employees, by employees, loan of employees, right of employees, employees income, claim of employees, employees development, employees' pension]
4	Phrases in previous line of number of employee	[number of employees, number of permanent employees, company, company and child entity, child company, each employing as many as, composition of the committee, permanent employees]
5	Phrases in next line of number of employee	[permanent employees, employee, each employee, audited, unaudited, unaudited, (unaudited), people, none]

Based on the value of identified feature sets, we grouped target information into two groups. First group consists of name of company, period of document and currency, which have clear orthographic and layout features. This target information has good regularity and included in the most frequent information. Second group consists of revenue and number of employee, which do not have specific orthographic, layout features but have rich of contextual, and list look-up features. The difference of characteristics owned induces the difference of strategy applied. We designed a multi-strategy approach to extract all target information. We separated these information and applied different strategy from beginning before designing extraction techniques. First strategy is developing extraction techniques using rule-based extraction method for name of company, period of document, and currency. Second strategy is applying machine learning-based extraction techniques for revenue and number of employee. Rule-based extraction techniques are developed using rule patterns. Rule patterns are defined by combining values of orthographic, layout, and contextual features owned by each target information. We constructed rule patterns in regular expression form. Here are rule patterns for target information.

Table 3. Rule Patterns for Name of Company Extraction

No	Pattern Rules for Name of Company Extraction
1	(combination {string = PT}) [(optional {symbol = . }) ({space}) ({Orthography type = word} {2-6}) : *slot* (optional {symbol = , })] :*name of company* ({space}) (combination {string = Tbk})

Table 4. Rule Patterns for Period of Documents Extraction

No	Pattern Rules for Period of Documents Extraction
1	[({Orthography type = digit} {2}) ({space}) ({string != and}) ({string != AND}) ({Orthography type = word}{1}) (optional{symbol = , }) ({space}) ({Orthography type = digit} {4})] : *period of document*

Table 5. Rule Patterns for Currency Extraction

No	Pattern Rules for Currency Extraction
1	(combination {string = presented in}) ({space}) [(optional {Orthography type = word} {1-3}) (optional {symbol = , }) (optional {symbol = $ }) (optional {symbol = \ }) ({space}) (optional {Orthography type = word} {1-4})] : *currency*
2	(combination {string = expressed in}) ({space}) [(optional {Orthography type = word} {1-3}) (optional {symbol = , }) (optional {symbol = $ }) (optional {symbol = \ }) ({space}) (optional {Orthography type = word} {1-4})] : *currency*

Second strategy is conducting extraction with classification approach. We used machine learning-based extraction method to construct extraction techniques. Developing extraction techniques are started from providing two data sets for training two classification models. We constructed two data sets from extracted features sets. First data set consist of extracted features sets of revenue tokens as positive class and extracted features sets of non-revenue numeric tokens as negative class. Second data set consists of extracted features sets of number of employee tokens as positive class and extracted features sets of numeric tokens (non-number of employee tokens) as negative class. We extracted thirteen feature sets, which are grouped into four feature groups. We extracted eleven features for number of employee tokens, which are also grouped into four feature groups. Four groups of revenue token's features are:

a. Orthographic feature group consist of:
    1. PrevTokenIsNumeric feature (*true* if one previous token is numeric and otherwise);
    2. NextTokenIsNumeric feature (*true* if one next token is numeric and otherwise).
b. Contextual feature group consists of:
    1. PrevOneToken feature (*string* one previous token);
    2. PrevTwoToken feature (*string* two previous token);
    3. NextOneToken feature (*string* one next token);
    4. NextTwoToken feature (*string* two next token);
    5. PrevLine feature (*string* sentence in one previous line);
    6. NextLine feature (*string* sentence in one next line).
c. List look-up features group consists of:
    1. ContainBoU feature (*true* if one previous token is part of unigrams as revenue identifier);
    2. ContainBoB feature (*true* if two previous token is part of bigrams as revenue identifier);
    3. NotContainNonBoU feature (*true* if one previous token is not part of unigrams as non-revenue identifier);
    4. NotContainNonBoB feature (*true* if two previous token is not part of bigrams as non-revenue identifier).
d. Layout feature is TokenOrderInLine feature (numeric that state position of token in line).

Number of employee tokens has same orthographic, contextual, and layout feature groups with feature groups of revenue tokens. The difference is on list look-up feature group. Here are list look-up feature group for revenue tokens:

1. ContainBoPW feature (*true* if one or two previous token is part of previous phrases of number of employee);
2. ContainBoNW feature (*true* if one or two next token is part of next phrases of number of employee);
3. ContainNonBoB feature (*true* if one previous token is not part of bigrams as non-number of employee identifier);
4. ContainBoWPL feature (*true* if tokens in previous line are part of phrases in previous line of number of employee);
5. ContainBoWNL feature (*true* if tokens in next line are part of phrases in next line of number of employee);

We constructed classification models using Naïve Bayes and Support Vector Machines (SVMs) algorithms. Naïve Bayes is a competitive algorithm for classification task although its constructed model is simple and the existence of independence assumption. Naïve Bayes classifier has been constructed for sentiment analysis and succeeds to reach classification accuracy 88.80% [8]. While F. Peng et al [9] and Kibriya et al [10] have improved Chain Augmented Naïve Bayes and several kinds of Multinomial Naïve Bayes for text classification. SVMs algorithm is also proved in classification task. Pang et al [11] conducted movie review classification using Naïve Bayes, Maximum Entropy and SVMs. They concluded that SVMs performed better than Naïve Bayes did although the difference of performance is not significant. Takeuchi and Collier [12] explored SVMs for named entity recognition in biology domain while Kudo and Matsumoto [13] introduced a SVMs-based text-chunking framework. Two constructed classification models are tested to know its performance. Classification model with highest performance is used for extraction task.

IV. Experiments and Results

We have conducted some experiments to get the best result. We used 939 financial report documents and started developing information extraction techniques from preprocessing data. Each target information on all documents are annotated using natural language processing toolkit. While annotation process, we found that some documents do not present revenue and number of employee information. Annotated documents are tokenized into paragraph and row collections. Then, row collections are tokenized again into list of numeric tokens. We only considered numeric tokens in order to reduce the size of data sets, to speed up the process of computing and to save the memory. We also analyzed content and structure of each document to get the characteristics of target information. We identified orthographic, contextual, layout, and list look-up features. Preprocessing data resulted 8.680.133 lines where each document contains 1.000-15.000 lines. Rule patterns are applied in every line. If text in line match with rule patterns, it means that text contain target information. Here are results of rule-based extraction:

Table 6. Results of Rule-based Extraction Method

Performance	Name of Company	Currency	Period of document
Number of tokens	939	939	939
Number of extracted tokens	939	939	939
Number of accurate extracted tokens	935	932	925
Number of not accurate extracted tokens	4	5	14
Recall	0.996	0.993	0.985
Precision	0.996	0.993	0.985
F1	0.996	0.993	0.985

Table 6 tells us that developed rule-based extraction technique is successful in extracting name of company, period of document and currency tokens. This technique gains precision, recall, and F1 more than 0.98. It means that rule patterns can extract tokens and well representative. It is successful because tokens have clear orthographic features. Failure is caused by inconsistent data presented in some documents, such as a document have two different names of company or two of document periods.

Machine learning-based extraction method is used to build extraction techniques for extracting revenue and number of employee information. We built two extraction techniques for information using two data sets. First data set is for building revenue extraction technique, which consists of 2,755 positive classes and 148.830 negative classes. Second data set is for building number of employee extraction technique that consists of 979 positive classes and 14,460 negative classes. Here are results of machine learning-based extraction for revenue and number of employee tokens:

Table 7. Results of Machine Learning-based Extraction Method for Revenue Tokens

No	Feature Set Group	Recall	Precision	F1
1	List lookup feature group	0.664	0.974	0.790
2	Contextual feature group	0.981	0.996	0.989
3	List lookup and orthographic feature group	0.664	0.974	0.790
4	Contextual, orthographic, list look-up and layout feature group	**1**	**1**	**1**

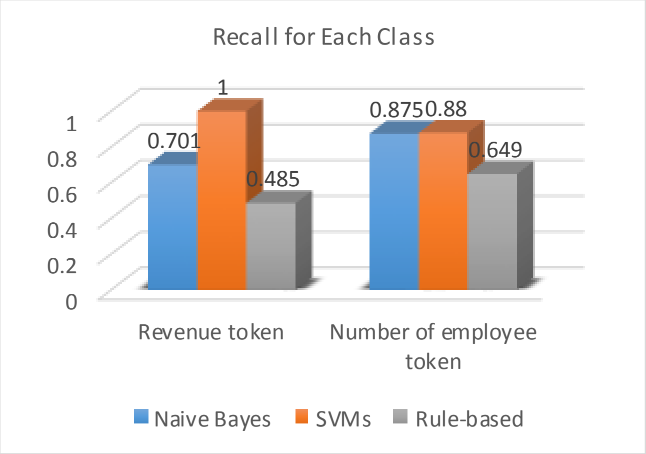
Table 7 informs us that list look-up features only can extract 66.4% revenue tokens but the extracted tokens are accurate by this type of feature. Orthographic feature addition does not give effect both of recall and precision. It is caused that orthographic feature for revenue token are not specific. Non-revenue numeric token is also preceded and followed by numeric token. Contextual feature group is successful to gain precision and recall more than 0.98. This success is caused contextual feature group can capture characteristics around revenue tokens. Use of all feature groups increases precision, recall, and F1. This combination is successful extracting all revenue tokens accurately. It is successful because this combination can capture all possible values of feature both feature of revenue token and feature of tokens around revenue token.

Table 8. Results of Machine Learning-based Extraction Method for Number of Employee Tokens

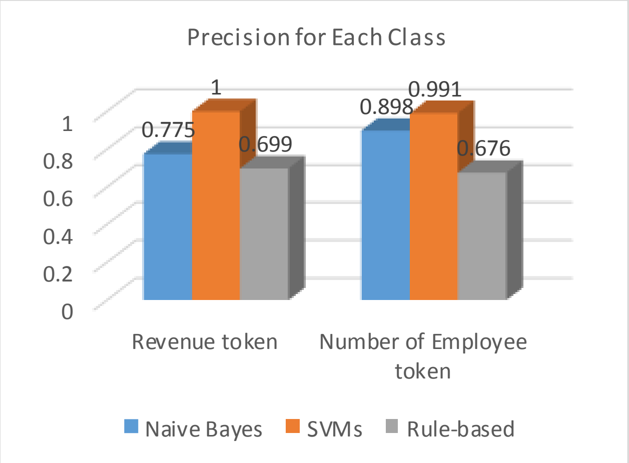
No	Feature Set Group	Recall	Precision	F1
1	List look-up feature group	0.929	0.899	0.914
2	Contextual feature group	0.872	0.990	0.927
3	List look-up and orthographic feature group	0.929	0.909	0.919
4	Contextual, orthographic, list look-up and layout feature group	**0.880**	**0.991**	**0.932**

Contrary to revenue tokens, list look-up features can extract more number of employee tokens and value of recall is higher than precision. Contextual feature group can gain F1 little higher than list look-up feature group. Contextual feature group can gain precision more than 99% because it can capture all possible values of feature both feature of number of employee tokens and feature of tokens around number of employee tokens. Combination of all feature groups can increase both recall and precision value of extraction. After analyzing the extraction results from each feature group, we explored machine learning algorithm to know performance of extraction. We explored Naive Bayes and Support Vector Machines (SVMs) then we compared its results with results of rule-based extraction.

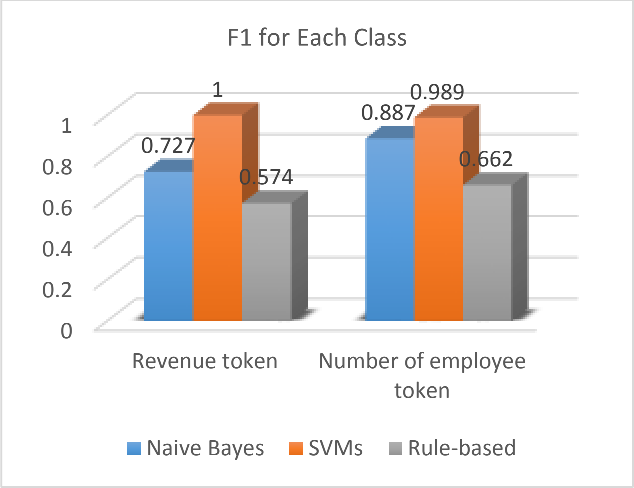
Graphic 1. Recall for Each Class



Graphic 2. Precision for Each Class



Graphic 3. F1-Measures for Each Class



Of two algorithms were tested on the best feature group (contextual, orthographic, list look-up and layout feature group), it is known that the algorithm of Support Vector Machines (SVMs) are able to perform better the learning process. It can be seen from the values of precision, recall, and F1 produced by SVMs algorithm. Performance of SVMs is highest compared with Naïve Bayes and rule-based extraction technique for revenues and number of employees classes. SVMs are able linear separating positive class and negative class in each training data.

V. Conclusion and Future Works

We have constructed accurate information extraction techniques in financial domain. These techniques are able extracting name of company, currency, period of document, revenue and number of employee information from financial report documents. Different with other works, we applied a multi-strategy approach in constructing extraction techniques. We assumed that the difference of characteristics owned by each target information, needs different strategy. Our assumption is proved by experiments. First strategy is applying rule-based extraction method on target information, which have good regularity on orthographic and layout features. Second strategy is applying machine learning-based extraction method on target information, which are rich of contextual and list look-up features.

We defined rule patterns by combining values of orthographic, layout, and contextual features. Rule patterns succeed to extract each name of company, period of document and currency token with high performance. It succeeds to reach F1-measure more than 0.98. On machine learning-based extraction method, we trained classification models using extracted feature groups, which formed positive and negative classes. Then, we experimented to extract revenue and number of employee tokens using best classification model. Combination of feature groups have an affect on performance of extraction. Contextual and list look-up features can be good identifier for revenue and number of employees tokens. Value of two these features can also distinguish revenue and number of employees tokens with other numeric tokens.

Extraction techniques developed by SVMs learning algorithms produce better extraction performance when compared with the Naïve Bayes algorithm. SVMs algorithm is able to capture the diversity or variation of the features of the token revenue and number of employees. It is not recommended to use rule-based extraction methods on two these target information. Extraction techniques are built with rule-based methods are not significant because the rules should be clearly defined and include all the features of value. If the rules are defined very detailed, it tends to produce a high precision but low recall. If the rules are defined global, it tends to produce a high recall but low precision. From these results, suggestions for further researches as follows: much more financial information extracted, assessing deeper learning algorithms and assessing the parameters used by each algorithm in building classification models.

References

[1] J.-L. Seng and J. T. Lai, An Intelligent Information Segmentation Approach to Extract Financial Data for Business Valuation, *Expert Syst. Appl.*, vol.37, no. 9, pp. 6515–6530, Sep. 2010.

[2] Badan Pusat Statistik, *Indeks Tendensi Bisnis dan Indeks Tendensi Konsumen*. 2013, p. 100.

[3] S. Sarawagi, Information Extraction, *Found. Trends Databases*, vol.1, no. 3, pp. 261–377, 2008.

[4] H. H. Malik, V. S. Bhardwaj, and H. Fiorletta, Accurate Information Extraction for Quantitative Financial Events, in *The 20th ACM International Conference on Information and Knowledge Management*, 2011, pp. 2497–2500.

[5] P. Andre and S. Ratte, "Classifier-based Acronym Extraction for Business Documents," *Knowl. Inf. Syst.*, vol. 29, no. 2, pp. 305–334, 2011.

[6] M. Sheikh and S. Conlon, A Rule-based System to Extract Financial Information, *J. Comput. Inf. Syst.*, vol.10, p. 10, 2012.

[7] H. Han, C. L. Giles, E. Manavoglu, H. Zha, and E. A. Fox, Automatic Document Metadata Extraction using Support Vector Machines, in *2003 Joint Conference on Digital Libraries*, 2003, pp. 37–48.

[8] V. Narayanan, I. Arora, and A. Bhatia, Fast and Accurate Sentiment Classification Using An Enhanced Naive Bayes Model, in *Intelligent Data Enginnering and Automated Learning*, 2013, pp. 1–8.

[9] F. Peng, D. Schuurmans, and S. Wang, Augmenting Naive Bayes Classifiers with Statistical Language Models, *Inf. Retr. Boston.*, **7**, no. 3–4, pp. 317–345, 2004.

[10] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, Multinomial Naive Bayes for Text Categorization Revisited, *Adv. Artif. Intell.*, **3339**, pp. 488–4999, 2004.

[11] B. Pang, L. Lee, H. Rd, and S. Jose, Thumbs up ? Sentiment Classification using Machine Learning Techniques, in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 1988, pp. 79–86.

[12] K. Takeuchi and N. Collier, Use of Support Vector Machines in Extended Named Entity Recognition, in *COLING-02, The 6th Conference on Natural Language Learning*, 2002, pp. 1–7.

[13] T. Kudo and Y. Matsumoto, Chunking with Support Vector Machines, in *NAACL '01 The 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technology*, 2001, **816**, pp. 1–8.