A comparative study on ML-based approaches for Main Entity Detection in Financial Reports

Thanos Konstantinidis, Yao Lei Xu, Tony G. Constantinides, Danilo P. Mandic
Department of Electrical and Electronic Engineering
Imperial College London
London, United Kingdom
Email: {a.konstantinidis16, yao.xu15, a.constantinides, d.mandic}@imperial.ac.uk

***Abstract*—Modern AI technologies which exploit the classification and/or prediction capacities of Deep Neural Architectures demonstrate superior performance to traditional approaches in most cases. However, they come with the unavoidable shortcoming of lack of transparency in their outcomes. This attribute renders them unsuitable for big industrial sectors, such as finance, investment management, etc. Specifically, their "black-box" nature makes them unattractive in cases where human understanding in the decision making process is required and may be legally mandatory. In such cases, traditional (i.e., non-deep learning) ML approaches are still preferred, to minimize for example the presence of false positives. In this context, this paper introduces an unsupervised, trustful, bottom-up probabilistic approach for Named Entity Recognition (NER) in financial reports, while in parallel it provides a comparative study on well-known ML-approaches in terms of their performance. The proposed approach builds on the probability of appearance of representative tokens within the given reports and utilizes Kronecker's Delta and the Total Probability Theorem to construct a probabilistic model that estimates the overall classification probability of a document.**

I. Introduction

Supervised learning methods have been used for text classification for decades achieving outstanding results in performance measurements. In order to achieve a high-performance score, standard algorithms, like Logistic Regression, Linear Support Vector Machines and Gaussian Naïve Bayes require a well-defined training set that will estimate their initial parameters, decrease their bias and prevent overfitting. Various methods have been proposed on how to evaluate successfully a classification algorithm, a widely used one is the Cross-Validation method, also used in this current work. As it will become apparent below, a text classification needs a well-designed pre-process analysis to extract the main characteristics that describe a document and which are then used to the classifiers. Choosing an inefficient training set may result in misclassification issues and low accuracy metrics. In practice, if it is tractable, we should avoid the training process of a classification model and make predictions without estimating initial parameters of the classifiers. In the current work we propose a probabilistic framework for entity detection which avoids the above-mentioned training process and deals directly with the estimated probability distribution density functions of the extracted attributes (tokens/n-grams) of a document to make classification prediction for a set of entities. The entities are well defined and expressed as a specific set of unique tokens/n-grams that represent each entity. Finally, we develop a probabilistic method based on the combination of two metrics

- the probability distribution functions of the tokens in a document, and
- the tokens describing each entity

That framework estimates the classification probability of a single document to a set of matching entities, where the highest classification probability corresponds to the matching predicted entity.

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

II. Related Work

A. Entity Recognition in Natural Language Processing

The term Named Entity Recognition (NER) was introduced for the first time in 6th Message Understanding Conference (MUC-6) [1]. A named entity is a token (aka. word or set of words) or a set of tokens representing a real-word entity [2]. NER is a sequence labelling problem and each observation, which has already been detected as entity, is assigned to a certain tag [2]. The entities' tags which were introduced for the first time in [1] was person, organization and location. In text mining research it is essential to identify such entities.

The process of entities' identification is defined as Named Entity Recognition (NER). This process is a subtask [3] or an assist task [4] of Information Extraction and aims at identifying semantics in unstructured data [6] and extracting respective components [5]. NER as a methodology faces two challenges: first the robust and optimized identification of named entities in a text and second the classification of these entities among predefined categories [7].

One basic problem that has puzzled researchers is the definition of the text boundaries within which, each entity is referenced. For this purpose, researches in domain, divided the text in chunks [8]. Chunks are noun phrases that are not recursive within a given document, and use the IOB format for the definition of entity boundaries. In each chunk we can assign three labels:

1) 'I' means that an entity tag is inside the chunk,
2) 'O' means that an entity tag is outside the chunk
3) 'B' means that an entity tag is in the beginning of the chunk.

Another method for boundaries recognition is TBL which analyses the form of a word (capital letters, lowercase letters, Capitalization of first digit) and the correspondence of a word with words in a dictionary [9]. It uses Hidden Markov Model methods to identify an entity and classify that entity in one of predefined categories. In this process, it is essential to annotate the data [13].

Named Entity Recognition contributes to question answering, summarization and machine translation [11]. Question answering (QA) is the task of automatically finding the best matching document (among a corpus of documents) to a given question. The type of entities included in a document is the criterion of rejecting or not this document as a possible answer. For example, if we want answers about financial questions, the documents which describe entities related to locations will be rejected [4]. Moreover, if multi-labels are assigned in an entity, the recall of the question answering process is improved (with simultaneous noise increase) [12]. Recent research makes it possible to reduce the noise level in QA [13] and to improve the quality in data selection against the quantity of data [14].

[11] and [15] create a multilingual machine translation framework by using the respective web links for multiple languages in Wikipedia. These links are useful for the projection of each entity into other languages [3][11].

Another Natural language Processing (NLP) task is summarization. In this process we examine the contribution of each sentence in the general meaning of text [17][18]. Sentences containing named entities are considered more important than other sentences of the document for the extraction of document's subject [16].

The classification problem of named entities, which is aforementioned, is related with sentiment analysis according to recent research [20]. More specifically, classification can be improved by assigning an opinion tag in each detected entity [19] and via classification of sentence type using BiLSTM-CRF and Recurrent Neural Networks (RNN) machine learning algorithms [21]. State-of-the-art research makes it possible to replace domain-specific lexicons with an automatic generated dictionary using deep learning algorithms [7].

B. Entity Recognition in Business data

Entity recognition contributes also in business research. More specifically, it plays a crucial role in product classification and recognition [22][25], in detection of novelty business terms and articles [28] and in reducing the noise generating processes of heterogeneous text sources e.g. Twitter posts or financial articles [29]. For example entities (companies) combined with other entities (products) and the related sentiment would be very useful for decision making in the field of business administration [31]. Businesses can use such applications to create products' attribute-value pairs or to find similarities between products. In addition, new documents can be used e.g. Twitter texts can be processed and their sentiment analyzed. The advantages of named entity recognition can also help in economics, where finance related text or articles can be decomposed into company features that describe the financial state of the company, thus supporting the investment process.

Product mentions are noun phrases, which include at least one common noun or proper noun related to a product [23]. Although, in the first steps of named entity recognition, entities are separated in three semantic categories –person, organization and locality [1], recent research enriches or changed these predefined categories. More specifically, [29] assigns the entities in four categories: person, organization, hardware, software. In addition, [30] divided entities among predefined categories which tagged as date, organization, location, money, percentage, person, or time while [31] follows the initial approach of [1].

Numerous approaches ended up with promising results. In the field of product-entity identification, [26] created an extensive dataset with products' offers and use the named entity recognition (NER) technique to pair products with the highest similarity. [24] achieved the same result after utilizing the product description to create attribute- value pairs. It was found that attribute-value pairs are more useful than atomic entities in fulfilment of missing value attributes. [25] aimed at improving the quality of text annotations in attribute-value extraction with the assistance of KB databases and an unsupervised learning method. In general, product mention detection, is very difficult due to the inconvenience of defining the boundaries of a noun phrase [26]. In [23] the CompanyProvidesProduct approach is proposed using combinations of predefined lexical patterns to identify the product mentions.

There is no extensive and published research addressing the problem of novel entity recognition in products and news' posts. Specifically, [27] use eBay product lists to extract known brands with typographical errors or novel brands. For this purpose, they use named entity recognition with combined bootstrapping statistic methods. In [28] a novel detection system is proposed using textual posts of financial blogs, based on only the textual content. The articles were separated into four categories –product, company, marketing, and finance-. This system gives the ability to businesses to find opportunities by the analysis of textual data. [29] analyzed Twitter data and retained only tweets which included named entities related to companies. [30] finds that stock market prediction achieves better results by using only the proper nouns of texts rather than using the named entities or entire noun phrases. This research refutes all other researches that utilize named entity recognition (NER) techniques.

III. CONTRIBUTION

The probabilistic framework proposed in the current work contributes to the unsupervised learning predictive algorithms domain. It has the ability to produce estimations that only rely on the probability distribution of an n-gram that belongs to a specific document: report, review etc. The above-mentioned ability differs from the well-known supervised learning algorithms e.g. Logistic Regression, Support Vector Machines etc., which need an efficient training and validation set to train the model and so to estimate their initial parameters. Herein, the predictive rate is strongly associated with the efficient extraction of the tokens that exist in a text body e.g. document, tweet etc., and mapping those tokens to an entity/category. In essence we map the tokens found in a document to the tokens describing the entity. Thus we classify that text body as related to an entity, assigning a probability score to that relationship. The more accurate and efficient the n-gram selection, the more accurate the classification.

IV. Methodology

The proposed method has two stages (1) Create a vector of tokens per entity (dictionary). This step can be manual, manually assisted or supervised; (2) Match the tokens of a report to the dictionaries of candidate entities. This step is unsupervised.

A token is an n-gram [1-W words, we used W $\leq$ 3], it can be more accurate than just words, e.g. 'Steve Jobs' is different and clearly has a stronger relationship with Apple than either 'Steve' or 'Jobs'.

Let us define the incoming text report as r. For reasons of simplicity, any information encoded in the grammar or the syntax of the incoming text r will be ignored, and r will only be treated as a set of tokens $r \equiv T_r = \{token_1 \quad token_2, \ldots, token_M\} = \{t_1 \quad t_2, \ldots, t_M\}$. So a report has m unique tokens and M tokens in total.

We assume that each financial document r refers to and reports about a main entity (i.e. a certain company, a certain stock market asset, etc.) and expands about it, so we can assign this information to entity ID. For the rest of this paper we will use the term portfolio asset and company interchangeably, as an asset belongs to a company. In particular, ID is defined as a vector, containing the set of the most relevant tokens $ID \equiv T_{id} = \{token_1 \quad token_2, \ldots, tokenL\} = \{t_1 \quad t_2, \ldots, t_L\}$, where L is the number of unique tokens. In real-world application scenarios, the vector may be algorithmically extracted from a pre-annotated set of documents and/or manually enhanced/defined by an expert (e.g. the portfolio manager himself, etc.). The challenge is to apply the correct match-making between two sets T_r and Tid or better expressed to estimate the probability $p(ID \vee r)$.

A. Dictionary Generation

Provided a portfolio of K assets (i.e. IDs) and the total set of N pre-annotated documents, we need to construct a small, but indicative word vectors (a.k.a dictionary) for each ID, e.g. $ID_{Apple} = \{\text{iphone, ipad, macbook,' Steve Jobs'}, \ldots\}$. Three commonly used metrics from Information Theory can been used, i.e.

(1) the information gain IG, which measures the number of bits of information obtained for category prediction by knowing the presence or absence of a token$_l$ in a document r.

$$IG(t_l, ID_k) = \sum_{ID \in \{ID_k, (ID)'_k\}} \sum_{t \in \{t_l, t'_l\}} P(t, ID) \log \frac{P(t, ID)}{P(t) \cdot p(ID)}$$

(2) the chi-square χ^2, Chi-square measures the lack of independence between a term t and a category c_i and can be compared to the chi-square distribution with one degree of freedom to judge extremeness.

$$\chi^2(t_l, ID_k) = \frac{\left[N \left(P(t_l, ID_k)P(t'_l, (ID)'_k) - P(t_l, (ID)'_k)P(t'_l, ID_k)\right)\right]^2}{P(t_l)P(t'_l)p(ID_k)p((ID)'_k)}$$

(3) the mutual entropy MI, which is a measure of dependence between variable (a term token$_l$ and a category ID), if MI for token$_l$ is zero then a term token$_l$ and a category ID are independent.

$$MI(t_l, ID_k) = \log \frac{P(t_l, ID_k)}{P(t_l) \cdot p(ID_k)}$$

By combining these aforementioned scores, the word vectors are generated, which contain the most representative words for each ID. Of course, as stated, in practical applications, these vectors may be manually refined by the portfolio managers themselves.

B. Classification Probability Estimation

In order to classify an incoming report to a certain asset of the portfolio, a probabilistic approach is followed. First we have to introduce Iverson's Bracket, also known as Kronecker's Delta generalization, in order to proceed in further calculations.

For a statement P, we define it as

$$[\![P]\!] = \begin{cases} 1 & \text{if } P \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

The above function can also been written as,

$$[\![\text{P is True}]\!] = 1 \tag{1}$$

As stated above we wish to assign asset ID to a report r. Let r be a set of tokens $r = \{token_1, \ldots, token_M\}$, having M tokens. For simplifying the calculations we set $token_i = t_i, \forall i \in \{1, \ldots, M\}$. We denote as m the number of unique tokens in the report r and set as $N^r_{t^*_i}$ the count of occurrence of token t^*_i in r. Clearly, $\sum_{i=1}^{m} N^r_{t^*_i} = M$. We denote the estimation of occurrence probability of a certain token t_i in r as the maximum likehood probability estimation $\Pr(t^*_i | t^*_i \in r) = \frac{N^r_{t^*_i}}{\sum_{i=1}^{m} N^r_{t^*_i}}$, which stands for the equation $\sum_{i=1}^{m} \Pr(t^*_i | t^*_i \in r) = 1$. A unique probability space $(\Omega, \mathcal{F}, P)$ is thereby created, where as Ω we denote the sample space of all possible outcomes, $\mathcal{F}$ the set of events and P the probability measure of the events. Translating the above statement in our case, we set as $\Omega \equiv r$, $\mathcal{F} = \{t^*_1, \ldots, t^*_m\}$ the set of unique tokens t^*_i in r and as $P = \{\Pr(t^*_i | t^*_i \in \mathcal{F})\}$.

We have denoted above an asset of the portfolio ID as $ID = \{\sigma_1, \ldots, \sigma_L\}$ to be represented as a set of L unique tokens σ_i. The probability of the report r matching asset ID is calculated as the probability of a token σ_i to belong in the probability space $(\Omega, \mathcal{F}, P)$. In order to accomplish this we follow the Law of Total Probability for each asset ID for the probability space $(\Omega, \mathcal{F}, P)$ of each report r as,

$$\Pr(ID) := \Pr(ID \cap r | \mathcal{F}) = \sum_{i=1}^{m} \Pr(ID, t^*_i \in \mathcal{F}) \tag{2}$$

Since the tokens $t^*_i \in F$ are assumed independent, equation (2) holds. We expand the above equation as,

$$\begin{aligned} \Pr(ID) &= \sum_{i=1}^{m} \Pr(ID, t^*_i \in F) \\ &= \sum_{i=1}^{m} \Pr(ID | t^*_i \in F) \cdot \Pr(t^*_i \in F) \\ &= \sum_{i=1}^{m} \Pr(ID | t^*_i \in F) \cdot \frac{N^r_{t^*_i}}{\sum_{i=1}^{m} N^r_{t^*_i}} \end{aligned} \tag{3}$$

It is observed in equation (3) that there is a term, $\Pr(ID|t_i^* \in F)$, which needs to be defined appropriately in order to calculate probability $\Pr(ID)$. As already mentioned each asset ID is defined as a set of unique tokens $\{\sigma_1, \ldots, \sigma_L\}$. We expect that in order to classify a report r in an asset ID, tokens from r have to be matching with the tokens in ID and vice versa. In fact, we set the conditional probability $\Pr(ID|t_i^* \in F)$ to be equal to,

$$\Pr(ID|t_i^* \in F) = [t_i^* \in ID|t_i^* \in F] \tag{4}$$

where $[t_i^* \in ID|t_i^* \in F]$ is the Iverson's bracket for the statement $P = \{t_i^* \in ID|t_i^* \in F\}$ which holds for equation (1). In mathematical terms, we denote that $t_i^* \in ID$ holds if and only if $\exists j \in \{1, \ldots, L\} : t_i^* \equiv \sigma_j$ for $i \in \{1, \ldots, m\}$. Equation (4) can also be translated as an activation function for ID according to the tokens that coexist in both ID and r, for every asset ID and every report r.

The final classification probability will thereby held by the equation,

$$\Pr(ID) = \sum_{i=1}^{m} [t_i^* \in ID|t_i^* \in F] \cdot \frac{N_{t_i^*}^r}{\sum_{i=1}^m N_{t_i^*}^r} \tag{5}$$

The above probability takes values in the interval $[0, 1]$ with the perfect matching classification to be held if and only if $\Pr(ID) = 1$, which stands for $[t_i^* \in ID|t_i^* \in F] = 1, \forall t_i^* \in F$. There is a probability of a report r to not be classified in an asset ID if and only if $\Pr(ID) = 0$, which stands for $[t_i^* \notin ID|t_i^* \in F] = 1, \forall t_i^* \in F$ (in effect, there are not matching tokens between the report r and asset ID). In cases that there are more than 1 assets $ID_1, \ldots, ID_N$ which have a non-zero probability $\Pr(ID_j) > 0$ for all $j \in \{1, \ldots, N\}$, then the classification asset will be selected to be the one with the highest probability, $\Pr(ID_j) = \max_j\{\Pr(ID_j)|j = 1, \ldots, N\}$, amongst these.

If there are 2 or more probabilities with the maximum value among probabilities, then the classification is a random choice among them under a uniform probability distribution assumption $U(p)$, with $p = 1/\left(||\Pr(ID_j)|\Pr(ID_j) = \max_k\{\Pr(ID_k)|k = 1, \ldots, N\}||\right)$. Another choice is to classify the report as referencing more than one asset.

Finally, for each report r in order to make a classification assumption, first, we must calculate the probability for each asset ID_j, $j = 1, \ldots, N$, where N is the total number of discreet assets in the portfolio, according to the equation (5), and then choose the one with the highest probability to classify the report r. So, the predicted asset will be according to the following final equation,

$$ID \equiv \underset{ID_j}{\operatorname{argmax}} \left\{ ID_j : \sum_{i=1}^{m} [t_i^* \in ID|t_i^* \in F|] \left(\frac{N_{t_i^*}^r}{\sum_{i=1}^m N_{t_i^*}^r} \right) \middle| \forall j \right\} \tag{6}$$

V. Experimental Evaluation

A. Dataset

The dataset used consists consists of a collection of 7700 financial reports of Moody's Corporation, Moody's Investor Services Inc., Moody's Analytics Inc., and other licensors and affiliates (collectively "MOODY'S"). We used this set of reports to create the dictionaries for the entities, as described above. There are 184 unique assets (entities) and for every report there is only 1 matching entity. It has to be mentioned that this Dataset refers to full-text data and not a text consisting only of a small number of sentences (e.g. tweeter feeds).

B. Evaluation Methodology

The proposed framework was evaluated using the well-known performance curve, ROC Curve, The classification has been tested out of sample in a number of different proprietary datasets. These datasets are adequately large, consisting of thousands of text documents.

C. Evaluation Results

We evaluated and compared the results of the proposed framework with existing supervised learning classifiers such as Linear Support Vector Machine (LSVM), Multinomial Naïve Bayes (MNBayes), Random Forest Classifiers (RFC), Gradient Boosting Classifier (GBC) and Multinomial Logistic Regression (MLR). For the evaluation of the above-mentioned classifiers a 10-Cross-Validation method has been implemented. In order to choose the best matching parameters for the above models a grid search option has been applied. Each text was represented by a number of n-grams of varying length between $n = \{1, 2, 3\}$.

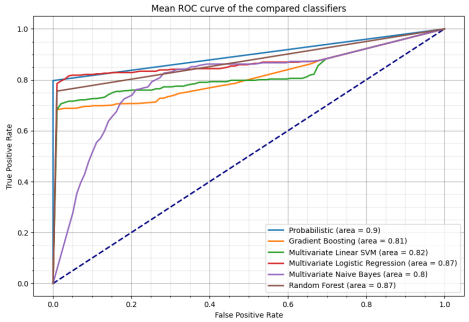

Fig. 1. ROC Curves Comparison of the evaluated models models. The results illustrate the average ROC curve from a 10-Cross-Validation experiment

Figure 1 illustrates the recognition capacity of the involved classifiers by the ROC Curves extracted represent the average of the classification results of a 10 Cross Validation process. The proposed approach demonstrates the largest Area under the Curve. A similar performance in their accuracy is presented in Table 1. All classifiers achieve an accuracy $> 80\%$. It is clear that the proposed approach achieves the highest accuracy rate (86%), despite it being an unsupervised process. Moreover, this is achieved while maintaining interpretability of the results.

TABLE I. Classification Algorithms and their accuracy for the processed texts according to their TF-IDF representation.

LSVM	MNBayes	RFC	GBC	MLR	Proposed
83%	85%	84%	81%	83%	86%

VI. Conclusion

A probabilistic approach for Named Entity Recognition (NER) in financial reports has been proposed. Experiment results demonstrate the advantages of the proposed framework, achieving better accuracy compared to other ML approaches.

REFERENCES

[1] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, vol. 1, 1996.

[2] C. C. Aggarwal and C. Zhai , Mining text data, Springer Science & Business Media, 2012.

[3] J. Daiber, M. Jakob , C. Hokamp and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in Proceedings of the 9th International Conference on Semantic System, ACM, 2013, pp. 121-124.

[4] D. Mollao, M. Van Zaanen and D. Smith, "Named entity recognition for question answering," 2006.

[5] B. Hachey and C. Grover, "Extractive summarisation of legal texts," Artificial Intelligence and Law, vol. 14, no. 4, pp. 305-345, 2006.

[6] M. Marrero, J. Urbano, S. Sanchez-Cuadrado, J. Morato and J. M. Gomez-Berbis, "Named entity recognition: fallacies, challenges and opportunities," Computer Standards & Interfaces, vol. 35, no. 5, pp. 482-489, 2013.

[7] E. Cambria, S. Poria, A. Gelbukh and M. Thelwall, "Sentiment analysis is a big suitcase," IEEE Intelligent Systems, vol. 32, no. 6, pp. 74-80, 2017.

[8] E. F. Sang and J. Veenstra, "Representing text chunks," in Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1999, pp. 173–79.

[9] R. Florian, "Named entity recognition as a house of cards: Classifier stacking," in proceedings of the 6th conference on Natural language learning-Volume 20, 2002, Association for Computational Linguistics, 2002, pp. 1-4.

[10] D. M. Bikel , R. Schwartz and R. M. Weischedel, "An algorithm that learns what's in a name," Machine learning, vol. 34, no. 1-3, pp. 211-231, 1999.

[11] J. Nothman, N. Ringland, W. Radford, T. Murphy and J. R. Curran , "Learning multilingual named entity recognition from Wikipedia," Artificial Intelligence, vol. 194, pp. 151-175, 2013.

[12] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT, Association for Computational Linguistics, 2003, pp. 1-8.

[13] E. Noguera, A. Toral, F. Llopis and . R. Munoz, "Reducing question answering input data using named entity recognition," in International Conference on Text, Speech and Dialogue, Springer, 2005, pp. 428-434.

[14] A. Toral, E. Noguera, F. Llopis and R. Munoz , "Improving question answering using named entity recognition," in International Conference on Application of Natural Language to Information Systems, Springer, 2005, pp. 181-191.

[15] A. E. Richman and P. Schone, "Mining wiki resources for multilingual named entity recognition," in Proceedings of ACL-08: HLT, 2008, pp. 1–9.

[16] V. Gupta and G. S. Lehal, "Named entity recognition for Punjabi language text summarization," International journal of computer applications, vol. 33, no. 3, pp. 28-32, 2011.

[17] B. Hachey and C. Grover, "Sequence modelling for sentence classification in a legal summarisation system," in Proceedings of the 2005 ACM symposium on Applied computing, ACM, 2005, pp. 292-296.

[18] A. Farzindar and G. Lapalme, "Legal text summarization by exploration of the thematic structure and argumentative roles," Text Summarization Branches Out, 2004.

[19] Y. Song, S. Jeong and H. Kim , "Semi-automatic construction of a named entity dictionary for entity-based sentiment analysis in social media," Multimedia Tools and Applications, vol. 76, no. 9, pp. 11319-11329, 2017.

[20] L. Gui, Y. Zhou , R. Xu, Y. He and Q. Lu, "Learning representations from heterogeneous network for sentiment classification of product reviews," Knowledge-Based Systems, vol. 124, pp. 34-45, 2017.

[21] . T. Chen, R. Xu, Y. He and X. Wang , "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," Expert Systems with Applications, vol. 72, pp. 221-230, 2017.

[22] Z. Wang, , X. Cui, , L. Gao, , Q. Yin, , L. Ke, and S. Zhang, , "A hybrid model of sentimental entity recognition on mobile social media," EURASIP Journal on Wireless Communications and Networking, vol. 2016, no. 2, p. 253, 2016.

[23] S. Schön, , V. Mironova, , A. Gabryszak, and L. Hennig, "A Corpus Study and Annotation Schema for Named Entity Recognition and Relation Extraction of Business Products," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

[24] R. Ghani, , K. Probst, Y. Liu, M. Krema and A. Fano, "Text mining for product attribute extraction," ACM SIGKDD Explorations Newsletter, vol. 8, no. 1, pp. 41-48, 2006.

[25] K. Shinzato and S. Sekine , "Unsupervised extraction of attributes and their values from product description," in Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013, pp. 1339-1347.

[26] H. Köpcke, A. Thor , S. Thomas and E. Rahm, "Tailoring entity resolution for matching product offers," in Proceedings of the 15th International Conference on Extending Database Technology, ACM, 2012, pp. 545-550.

[27] D. P. Putthividhya and J. Hu, "Bootstrapped named entity recognition for product attribute extraction," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1557-1567.

[28] H. Liang, , F. S. Tsai, and A. T. Kwee, , "Detecting novel business blogs," in 2009 7th International Conference on Information, Communications and Signal Processing (ICICS), IEEE, 2009, pp. 1-5.

[29] T. T. Vu, , S. Chang, , Q. T. Ha, and N. Collier, , "An experiment in integrating sentiment features for tech stock prediction in twitter," in Proceedings of the workshop on information extraction and entity analytics on social media data, 2012, pp. 23-38.

[30] R. P. Schumaker, and H. Chen, , "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," ACM Transactions on Information Systems (TOIS), vol. 27, no. 2, p. 12, 2009.

[31] S. P. Tripathi and H. Rai , "SimNER–An Accurate and Faster Algorithm for Named Entity Recognition," in 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), IEEE, 2018, pp. 115-119.