The FinSBD-2019 Shared Task: Sentence boundary detection in PDF Noisy text in the Financial Domain

Abderrahim Ait Azzi[1] , **Houda Bouamor**[2] , **Sira Ferradans**[1]
[1]Fortia Financial Solutions, France
[2]Carnegie Mellon University in Qatar, Qatar
name.surname@fortia.fr, hbouamor@cmu.edu

Abstract

In this paper, we present the results and findings of The FinSBD-2019 Shared Task on Sentence boundary detection in PDF Noisy text in the Financial Domain. This shared task was organized as part of The First Workshop on Financial Technology and Natural Language Processing (FinNLP), collocated with IJCAI-2019. The shared task aimed at collecting systems for extracting well segmented sentences from Financial prospectuses by detecting and marking their beginning and ending boundaries. The FinSBD shared task is the first to target the task of sentence boundary detection in the domain of Finance. A total of 9 teams from 7 countries participated in the shared task with a variety of systems and techniques.

1 Introduction

A vast amount of documents are constantly published online in machine-readable formats (generally PDF), containing not only text, but also other elements such as tables, images, and graphics. Therefore, most of the established PDF to text conversion products on the market (i.e., pdf2text) generate highly noisy unstructured texts containing abbreviations, non-standard words, false starts, missing punctuation, missing letter case information, and other text disfluencies. Building NLP applications customized for such texts is very challenging as most of the NLP tools (i.e. POS tagging, parsing, etc.) and applications (i.e. information extraction, machine translation) require as input a well-formatted clean text, where sentence boundaries are clearly marked [1].

Despite its important role in NLP, sentence boundary detection (SBD) has so far not received enough attention. Previous research in the area has been confined to formal texts only (news, European Parliament proceedings, etc.) where existing rule-based and machine learning approaches are extremely accurate (when the data is perfectly clean). No sentence boundary detection research to date has addressed the problem in noisy texts extracted automatically from machine-readable formats (generally PDF file format) files such as financial documents.

In this shared task, we focus on extracting well segmented sentences from Financial prospectuses by detecting and marking their beginning and ending boundaries. These are official PDF documents in which investment funds precisely describe their characteristics and investment modalities. The most important step of extracting any information from these files is to parse them to get noisy unstructured text, clean it, format information (by adding several tags) and finally, transform it into semi-structured text, where sentence boundaries are well marked.

In this paper we report the results and findings of the FinSBD-2019 shared task.[1] The Shared Task was organized as part of The First Workshop on Financial Technology and Natural Language Processing (FinNLP), collocated with IJCAI-2019.[2]

A total of 9 teams from 7 countries submitted runs and contributed 7 system description papers. All system description papers are included in the FinNLP workshop proceedings and cited in this report.

The large number of teams and submitted systems suggests that such shared tasks can indeed generate significant interest in the Finance and NLP research community.

2 Previous Work on SBD

While SBD is a foundational pre-processing task, previous research has been confined to clean texts in standard areas such as the news and limited datasets such as the WSJ corpus [2] or the Brown corpus [3]. SBD has been largely explored following several approaches that could be classified into three major classes: (a) rule-based SBD, using hand-crafted heuristics and lists [4]; (b) machine learning approaches to SBD [5; 6; 7; 3]; and more recently (c) deep learning methods [8]. Most of these approaches give fairly accurate results. These systems are based on a number of assumptions [4] that do not hold for noisy texts extracted automatically from PDFs (data is perfectly clean).

[1] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/shared-task-finsbd

[2] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/home

Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pages 74-80, Macao, China, August 12, 2019.

{'text': " UFF Sélection Alpha AINFORMATIONS CLÉS POUR L' INVESTISSEUR
Ce document fournit des informations essentielles aux investisseurs de cet OPCVM .
Il ne s' agit pas d' un document promotionnel . Les informations qu ' il contient vous
sont fournies conformément à une obligation légale , afin de vous aider à comprendre
en quoi consiste un investissement dans ce fonds et quels risques y sont associés . ..." ,

'begin_sentence': [8, 21, 31 , ...],

'end_sentence': [20, 30, 66, ...] }

Figure 1: Example of the data json file

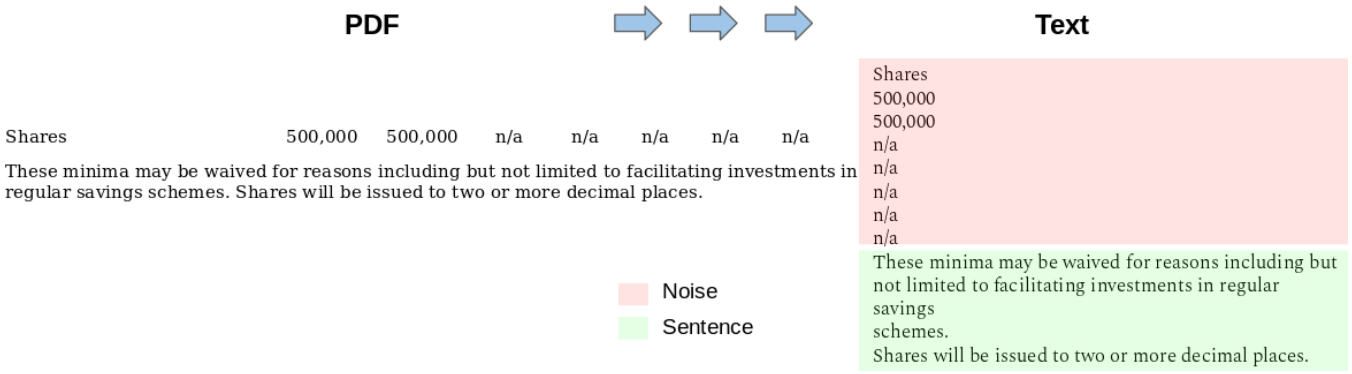

Figure 2: Example of a pdf to text conversion: Text extracted from a prospectus in a PDF format, spanning on several lines and with no punctuation marks (Target sentences are highlighted in green, Noise in Red)

Read et al., [1] conducted an analysis and review of commonly used *SBD* tools, but with a focus on generalization towards user-generated Web content. They evaluated several systems on a variety of data sets and report a performance decrease when moving from corpora with formal language to those that are less formal. Thus, designing and implementing approaches customized to different domains attracted the attention of several researchers. Griffis et al., [9] evaluated popular off-the-shelf NLP toolkits on the task of *SBD* for a set of corpora in the clinical domain. López and Pardo [10] tackle *SBD* on informal user-generated content such as web reviews, comments, and posts. Rudrapal et al,. [11] present a study on *SBD* in social media context. *SBD* from speech transcriptions has also gained a lot of attention due to the necessity of finding sentential segments in the stream of transcripts, automatically recognized [12; 13].

Although text extracted from financial documents such as financial prospectuses, faces problems of text quality caused by segmentation issues, SBD (similarly to most other NLP tasks) has not received much attention in this domain.

3 Task Description

As part of the First Workshop on FinTech and Natural Language Processing (FinNLP), we introduced the FinSBD shared task which aims at sentence boundary detection in noisy text extracted from financial prospectuses, in two languages: English and French. Systems participating in this shared task were given a set of textual documents extracted from pdf files, which are to be automatically segmented to extract a set of well delimited sentences (clean sentences). The data will be in a json format (i.e. figure 1) containing: **"text"**, that corresponds to the text to segment, **"begin_sentence"** and **"end_sentence"** correspond to all indexes of tokens marking the beginning and the end of well formed sentences in the text. It is important to note that the provided text is already segmented at the word level. All participants were asked to keep this segmentation since all tokens indexes are built based on it. The first token in the text has then the index 0 .

As stated in section 2, most of the previous research on sentence segmentation has been confined to clean texts in standard areas such as the news and limited datasets such as the WSJ corpus. However, the task of segmenting sentences extracted from noisy text, and more specifically text resulting from pdf conversion in the domain of finance is not much explored in the literature.

Figure 2 illustrates an example of a text extracted automatically from an English financial prospectus containing numerous issues ranging from missing punctuation to sentences spanning on several lines, in addition to the non-standard capitalization (very typical in financial texts).

Other issues are caused by the ambiguous use of full stop punctuation marks in several section numbers (i.e., "1.","2.") and to mark the end of a sentence. Also, the dash sign (-) could be used as a hyphen or to mark the math minus sign. Moreover, financial prospectuses contain a large number of financial institutions names that appear with their legal form abbreviations (i.e.,"S.A"for Société Anonyme,"LTD." for Limited Company, etc. Hence, applying commonly used sentence segmentation tools (i.e., Stanford sentence segmenter [14]) that typically rely on punctuation marks or capitalization in the sentence boundary detection (SBD) process is impractical.[3]

4 Shared Task Data

Next, we discuss the corpora used for the English and French subtasks.

4.1 Corpus annotation

Financial prospectuses are available online in a pdf format and are also made available from asset managers. We compiled a list of 11 prospectuses in English (140 pages on average) and 92 in French (25 pages on average). These prospectuses are first converted to a text document format using the freely available tool pdf2text[4]. Every line in these documents is tokenized at the word level. We extend the Keras tokenizer by adding several rule-based functions to take into account more cases (i.e., possessive form of words, acronym detection, etc.). We remove all non-ASCII characters resulting from the conversion step except the French accents.

We provided three bi-lingual (English and French) annotators with text files in both languages extracted automatically from financial prospectuses, along with their original PDFs.[5] We gave them detailed annotation guidelines and asked them to go through every text segment, understand it and mark the boundaries of what they estimate corresponds to a sentence. A sentence is defined as a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses. We deliberately asked them not to rely on capitalization and punctuation markers only. We use the online annotation framework BRAT [15]. The tool displays to the annotator each document in a text segment per line format.

The annotation labels used in building the corpus are the following:

- **Begin_Sentence** (**BS**): marks the first token of a sentence which can be a word, a character or bullet, etc.
- **End_sentence** (**ES**): denotes the token that comes at the end of a sentence whether it is a punctuation mark or a word in case the sentence does not end with a punctuation.

[3] `https://nlp.stanford.edu/software/tokenizer.html`

[4] `http://www.pdf2text.com/`

[5] This helps visualize the text in its original clean setup, which helps the annotators locate sentences more rapidly.

Annotation Challenges . The data annotation process was an arduous task due to:

1. the absence of the layout creates many ambiguities: some headers are transformed to lower-cased words and are put in the same lines as the sentences. For sentences that begin at the end of a PDF page and end in the next one, the footer content enters inside such sentences in the text file which makes it impossible to annotate them. So in this case, the solution was to remove manually these footers.
2. the absence of end of sentence punctuation markers. In fact, some sentences do not end with full stops e.g. "The DJ - UBS Index generally rolls the futures contract which is closest to expiry into the futures"
3. the punctuation errors appearing in some sentences, such as a period in the middle of a sentence e.g. *This Supplement forms part of the Prospectus dated 1 January 2015 for GAM Star (Lux) SICAV . and should be read in conjunction with that Prospectus.*
4. the excessive use of in-sentence lists mainly in English prospectuses. An in-sentence list is a sentence that contains a list of ordered sub-sentences usually using letter (a), (b) and so on, or numbers (1), (2) and so on.
5. the excessive use of Uppercase words that are neither proper nouns nor named entities (Shares, Class, Initial Subscription,..) which makes it less obvious to spot the beginning of a sentence.

4.2 Corpus Description

In the following, we provide an analysis of the data used for both subtasks: English and French.

In Table 1, #prospectuses indicates the number of prospectuses that were used in each data set; #Types is the total number of unique tokens in the text; and #Sentences is the number of segmented sentences in the text. % *OOV* words represents the rate of Out-of-Vocabulary words. We notice that the French Validation/Testing data contain higher *OOV* rate the English data ($\approx +5\%$).

In order to extend our analysis, we first report the percentage of sentences ending with a punctuation mark such as the full stop, column, and semi-column. Although this rate is higher than 93% for both language, it still shows that there are many sentences that do not contain any ending indicator as mentioned in the previous section notably in the french testing set ($\approx 7\%$). Then, we report the percentage of sentences that start with a capital letter. This percentage is around 85% for the English data, which means that in many cases, capitalization is not an indicator of the beginning of sentences, which shows that our task is more sophisticated than the traditional SBD tasks.

5 Participants and Systems

A total of 69 teams registered in the shared task, out of which 8 submitted a paper with the description of their method. The participants came from 7 different countries and belonged to 10 different institutions. The shared task was a success in bringing together private and public research institutions. As private, Accenture AI Labs, SeerNet Technologies LLC, OPT inc and AIG. As public, Heidelberg Institute for theoretical studies, University of Kyoto, Insight Center for Data Analytics (National University of Ireland Galway), the Hong Kong Polytechnic University and Harbin Institute of Technology (see Table 3 for more details).

In table 2, we show the details on the submissions per task. It is important to note that not all the participants that submitted a standard run, sent a paper describing their approach.

Participating teams explored and implemented a wide variety of techniques and features. In this section, we give a brief description of each system, more details could be found in the description papers appearing in the proceedings of the FinNLP 2019 Workshop.

Most participants formulated the problem as either a sequence-labeling task or as a word level classification task. In this context, the best performing methods are those that used word embeddings with a neural model mainly based on LSTMs, although other features were explored. Below is a short summary of each participating system. Teams Seernet and ISI do not appear because they did not send a descriptive paper.

AI_blues [19] In this work, the problem is stated as a sequence labeling task for which the participants use a CRF method. The input features for the CRF are mostly defined based on punctuation, lexical combinations of numbers and letters, presence of upper case letters, POS tags and some basic features (token length, is upper case, is lower case, token type).

NUIG [18] This team was the only one that took into account and explored the financial component of the task. They trained several character-level RNN embeddings using external financial text. These embeddings were used together with pretrained GLOVE [23] (for English) and FastText [24] (for French) embeddings. Finally, the system performs a sequence labeling using a BiLSTM-CRF model.

PolyU [22] The team defines a set of handcrafted features such as punctuation, presence of upper case, acronyms, and POS tags and train two models: (1) a multilayer neural network and (2) a random forest model. They show that cross-lingual training improves results for both languages.

mhirano [20] The features used in this system are pre-trained word2vec embeddings, POS tags, presence of capital letters, and alpha-numerical patterns. They serve as input to a multilayer perceptron trained to classify the central word of a given window. They also propose a second method that is rule-based and defined on sequences of token types.

HITS-SBD [21] This team proposed two methods: (1) random forest classifier on top of a TF-IDF representation

	English			French		
	Training	**Validation**	**Testing**	**Training**	**Validation**	**Testing**
# Prospectuses	9	1	1	74	9	9
# Tokens	904057	49859	56952	827852	119008	106577
# Types	13478	2926	3651	14267	6267	5610
# Sentences	22342	1384	1265	22636	3141	2981
% *OOV* words	—	10.93	13.75	—	15.9	19.22
% Punct. as end sentence	93.31	97.97	97.0	94.39	96.18	93.62
% Uppercase begin sentence	87.33	82.80	84.35	89.90	90.06	89.23

Table 1: Distribution of the Training, Validation and Testing sets used in the English and French corpora.

	# teams	**# std runs**
subtask EN	9	18
subtask FR	7	15
papers	8	-

Table 2: Statistics on the participation in the French and English subatsks.

of the word context, and (2) a ruled-based method based on pattern matching.

aiai [17] They defined the task as a classification task of the center word of a given window. They use two classification methods: (1) LSTM with attention and (2) CNN, both on top of pretrained Glove word level embeddings and specific word embeddings that encode upper case letters.

AIG [16] Similarly to many of the other teams, AIG implemented two models: (1) BI-LSTM with CRF on top of pretrained GLOVE word embeddings, and (2) a fined-tuned version of BERT for the sequence labeling.

6 Results and Discussion

In this section, we describe the evaluation metrics used in the shared task and we give an analysis of the results obtained for the various submitted systems.

Evaluation Metric Participating systems are ranked based on the macro-averaged F1 scores obtained on blind test sets (official metric). We also report the scores of **Begin_Sentence** (**BS**) and **End_Sentence** (**ES**), that are computed separately.

Table 4 reports the results obtained on FinSBD English by the teams detailed in the previous section. For the results on FinSBD French, please check table 5.

	English		
Team	**BS**	**ES**	**Average**
AIG1	0.88	0.89	**0.885**
seernet1	0.85	0.9	0.875
aiai1	0.83	0.91	0.87
isi1	0.83	0.89	0.86
NUIG1	0.81	0.9	0.855
isi2	0.82	0.89	0.855
AIG2	0.83	0.88	0.855
AI_Blues2	0.82	0.87	0.845
AI_Blues1	0.82	0.87	0.845
mhirano1	0.78	0.89	0.835
aiai2	0.79	0.88	0.835
NUIG2	0.81	0.85	0.83
HITS-SBD2	0.8	0.86	0.83
HITS-SBD1	0.8	0.86	0.83
PolyU_CBS-CFA_NN1	0.77	0.86	0.815
PolyU_CBS-CFA_RFC1	0.7	0.86	0.78
PolyU_CBS-CFA_RFC2	0.68	0.86	0.77
mhirano2	0.58	0.67	0.625

Table 4: Results obtained by the participants for the FinSBD English task. The teams are ordered by the F1 average value (last column).

	French		
Team	**BS**	**ES**	**Average**
seernet	0.91	0.93	**0.92**
aiai1	0.91	0.92	0.915
NUIG1	0.9	0.92	0.91
NUIG2	0.9	0.92	0.91
isi1	0.9	0.91	0.905
isi2	0.89	0.91	0.9
AI_Blues1	0.85	0.88	0.865
AI_Blues2	0.84	0.88	0.86
PolyU_CBS-CFA_RFC1	0.84	0.88	0.86
mhirano1	0.82	0.89	0.855
PolyU2	0.83	0.87	0.85
PolyU_CBS-CFA_NN1	0.83	0.87	0.85
PolyU_CBS-CFA_RFC2	0.81	0.88	0.845
mhirano2	0.67	0.68	0.675
aiai2	0.01	0.02	0.015

Table 5: Results obtained by the participants for the FinSBD French task. The teams are ordered by the F1 average value (last column).

Discussion Simple ruled-based methods based on obvious punctuation characters can perform very well on SBD, but in order to perform extremely well, we need to take into account the long tail exceptions specially

Team	Affiliation	Tasks
AIG [16]	American International Group, United Kingdom	English only
seernet	SeerNet Technologies, LLC, India	English and French
aiai [17]	OPT, Inc and Herbin institute of technology, Japan and China	English and French
isi	Information Sciences Institute (University of Southern California), USA	English and French
NUIG [18]	National University of Ireland Galway, Ireland	English and French
AI_Blues [19]	Accenture Solutions Pvt Ltd, India	English and French
mhirano [20]	The University of Tokyo, Japan	English and French
HITS-SBD [21]	Heidelberg Institute for Theoretical Studies, Germany	English only
PolyU_CBS [22]	The Hong Kong Polytechnic University, China	English and French

Table 3: List of the 9 teams that participated in Subtasks English and French of the FinSBD Shared Task.

present in noisy financial text extracted from pdf, which is the target corpus of this shared task. We can see this in the results. Two ruled-based methods, mhirano2 and HITS-SBD2, were proposed obtaining the 18th (0.625 F1 score in EN) and 13th position (0.83 F1) respectively. All the other methods implemented machine learning algorithms (AI_blues, HITS-SBD1, PolyU2) and deep learning methods (NUIG, PolyU1, mhirano1, aig, aiai). The best performing teams (NUIG1, aig1 and aiai1) implemented similar models: on top of GLOVE word embeddings a combination of (bi-)lstm with a CRF or attention layer.

Very little attention was payed to the fact that the corpus was from the financial domain. Only one team used financial features by training a language model on external financial text.

Finally, most participants understood how similar the task was to POS tagging and either used POS tags as features (PolyU, mhirano AI_Blues) or took inspiration from learning methods that performed well in POS tagging tasks (NUIG1).

7 Conclusions

In this paper we presented the setup and results for the FinSBD-2019 Shared Task on Sentence boundary detection in PDF Noisy text in the Financial Domain, organized as part of The First Workshop on Financial Technology and Natural Language Processing (FinNLP), collocated with IJCAI-2019. A total of 69 people registered and 9 teams from 7 countries participated in the shared task with a wide variety of techniques. The most successful methods were based on word embeddings (mostly GLOVE) followed by a (bi)lstm-crf (or an attention mechanism). The best average F1 score on the FinSBD French task was 0.92 and 0.885 for the FinSBD English.

We introduced a new data set on the SBD problem in text automatically extracted from PDF files for French and English. This scenario is very realistic in everyday applications which may explain the diversity of institutions that participated, from public universities to for profit organizations from the financial domain. In this sense, the shared task was a success since it was able to bring together researchers from different sectors.

Acknowledgments

We would like to thank our dedicated annotators who contributed to the building the French and English corpora used in this Shared Task: Anais Koptient, Aouataf Djillani, and Lidia Duarte.

References

[1] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[2] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[3] Dan Gillick. Sentence boundary detection and the problem with the us. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244. Association for Computational Linguistics, 2009.

[4] Gregory Grefenstette and Pasi Tapanainen. What is a word, what is a sentence?: problems of tokenisation. 1994.

[5] Michael D Riley. Some applications of tree-based modelling to speech and language. In *Proceedings of the workshop on Speech and Natural Language*, pages 339–352. Association for Computational Linguistics, 1989.

[6] Jeffrey C Reynar and Adwait Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*, pages 16–19. Association for Computational Linguistics, 1997.

[7] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[8] Marcos V Treviso, Christopher D Shulby, and Sandra M Aluisio. Evaluating word embeddings for sentence boundary detection in speech transcripts. *arXiv preprint arXiv:1708.04704*, 2017.

[9] Denis Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M Lai. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. *AMIA Summits on Translational Science Proceedings*, 2016:88, 2016.

[10] Roque López and Thiago AS Pardo. Experiments on sentence boundary detection in user-generated web content. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 227–237. Springer, 2015.

[11] Dwijen Rudrapal, Anupam Jamatia, Kunal Chakma, Amitava Das, and Björn Gambäck. Sentence boundary detection for social media text. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 254–260, 2015.

[12] González-Gallardo, Carlos-Emiliano, Torres-Moreno, and Juan-Manuel. Sentence boundary detection for french with subword-level information vectors and convolutional neural networks. *arXiv preprint arXiv:1802.04559*, 2018.

[13] Chenglin Xu, Lei Xie, and Xiong Xiao. A bidirectional lstm approach with word embeddings for sentence boundary detection. *Journal of Signal Processing Systems*, pages 1–13, 2017.

[14] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[15] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.

[16] Yan Huang Jinhua Du and Karo Moilanen. Aig at the finsbd task: Sentence boundary detection through sequence labelling and bert fine-tuning. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[17] Ke Tian and Zi Jun Peng. aiai at finsbd task: Sentence boundary detection in noisy texts from financial documents using deep attention model. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[18] Tobias Daudert and Sina Ahmadi. Nuig at the finsbd task: Sentence boundary detection for noisy financial pdfs in english and french. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[19] Ditty Mathew and Chinnappa Guggilla. Ai_blues at finsbd shared task: Crf-based sentence boundary detection in pdf noisy text in the financial domain. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[20] Kiyoshi Izumi Masanori Hirano, Hiroki Sakaji and Hiroyasu Matsushima. mhirano at the finsbd task: Pointwise prediction based on multi-layer perceptron for sentence boundary detection. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[21] Mehwish Fatima and Mark-Christoph Mueller. Hitsbd at the finsbd task: Machine learning vs. rule-based sentence boundary detection. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[22] Emmanuele Chersoni Natalia Klyueva Kathleen Ahrens Bin Miao David Broadstock Jian Kang Amos Yung Mingyu Wan, Rong Xiang and Chu-Ren Huang. Sentence boundary detection of financial data with domain knowledge enhancement and cross-lingual training. In *The First Workshop on Financial Technology and Natural Language Processing of IJCAI 2019*, 2019.

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[24] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.