

FinNLP 2020

**The Second Workshop on Financial Technology  
and Natural Language Processing  
in conjunction with IJCAI-PRICAI 2020**

**Proceedings of the Workshop**

January 5, 2021

Kyoto, Japan

# Sponsor



科技部臺灣大學人工智慧研究中心

科技部人工智慧技術暨全幅健康照護聯合研究中心  
MOST JOINT RESEARCH CENTER FOR AI TECHNOLOGY AND  
ALL VISTA HEALTHCARE

## Preface

It is our great pleasure to welcome you to the Second Workshop on Financial Technology and Natural Language Processing (FinNLP).

The aim of FinNLP is to provide a forum for international participants to share knowledge on applying NLP to the FinTech domain. With the sharing of the researchers in this workshop, we hope that the challenging problems of blending FinTech and NLP will be identified, the future research directions will be shaped, and the scope of this interdisciplinary research area will be broadened.

Due to a great trial of all of us, the COVID-19 virus, the IJCAI-PRICAI conference is postponed to January 2021. In order to accelerate the development of this field, we decide to publish the proceedings in advance. This year, the participants of FinNLP still bring several novel ideas to this forum. We also cooperate with Fortia Financial Solutions to hold two shared tasks in FinNLP, including learning semantic representations (FinSim) and sentence boundary detection in PDF noisy text (FinSBD-2).

We have many people to thank. Dialekti Valsmou Stanislawski and Ismail El Maarouf lead their teams to hold the successful shared tasks and help review several submissions. All of program committee members work very hard to provide their insightful comments to the submissions, and help us select the suitable papers for FinNLP-2020. Many thanks to all participants for submitting their interesting works and sharing their ideas. Besides, we would like to express our gratitude to the MOST Joint Research Center of AI Technology and All Vista Healthcare for financial support.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen  
FinNLP 2020 Organizers  
July 2020



# Organizations

## Organizers

Chung-Chi Chen, National Taiwan University

Hen-Hsen Huang, National Chengchi University

Hiroya Takamura, National Institute of Advanced Industrial Science and Technology

Hsin-Hsi Chen, National Taiwan University

## Shared Task Organizers

Willy Au

Abderrahim Ait Azzi

Bianca Chong

Ismail El Maarouf

Youness Mansar

Virginie Mouilleron

Dialekti Valsamou-Stanislawski

## Program Committee

Paulo Alves, Universidade Católica Portuguesa

Alexandra Balahur, European Commission's Joint Research Centre

Paul Buitelaar, Insight Centre for Data Analytics at NUIG

Damir Cavar, Indiana University

Pablo Duboue, Textualization Software Ltd.

Jinhua Du, American International Group, Inc.

Ismail El Maarouf, Fortia Financial Solutions

Sira Ferradans, Independent Researcher

Kiyoshi Izumi, The University of Tokyo

Changliang Li, Kingsoft Corporation

Nedim Lipka, Adobe Inc.

Hiroki Sakaji, The University of Tokyo

Dialekti Valsamou-Stanislawski, Fortia Financial Solutions

Chuan-Ju Wang, Academia Sinica

Annie T.T. Ying, EquitySim

Wlodek Zadrozny, University of North Carolina in Charlotte



## Table of Contents

Financial News Annotation by Weakly-Supervised Hierarchical Multi-label Learning Hang Jiang, Zhongchen Miao, Yuefeng Lin, Chenyu Wang, Mengjun Ni, Jian Gao, Jidong Lu and Guangwei Shi.....	1
Variations in Word Usage for the Financial Domain Syrielle Montariol, Alexandre Allauzen and Asanobu Kitamoto.....	8
Detecting Omissions of Risk Factors in Company Annual Reports Corentin Masson and Syrielle Montariol.....	15
A Semantic Approach to Financial Fundamentals Jiafeng Chen and Suproteem Sarkar.....	22
Learning Company Embeddings from Annual Reports for Fine-grained Industry Characterization Tomoki Ito, Jose Camacho Collados, Hiroki Sakaji and Steven Schockaert.....	27
Unsupervised Discovery of Firm-Level Variables in Earnings Call Transcript Embeddings Daniel Edmiston and Zihong Park.....	34
Using Extractive Lexicon-based Sentiment Analysis to Enhance Understanding of the Impact of Non-GAAP Measures in Financial Reporting Stacey Taylor and Vlado Keselj.....	40
FinSBD-2020: The 2nd Shared Task on Sentence Boundary Detection in Unstructured Text in the Financial Domain Willy Au, Bianca Chong, Abderrahim Ait Azzi and Dialekti Valsamou-Stanislawski.....	47
PublishInCovid19 at the FinSBD-2 Task: Sentence and List Extraction in Noisy PDF Text Using a Hybrid Deep Learning and Rule-Based Approach Janvijay Singh.....	55
ai ai at the FinSBD-2 Task: Sentence, list and Item Boundary Detection and Items classification of Financial Texts Using Data Augmentation and Attention Ke Tian, Hua Chen and Jie Yang.....	62
Daniel at the FinSBD-2 Task: Extracting List and Sentence Boundaries from PDF Documents, a model-driven approach to PDF document analysis Emmanuel Giguet and Gaël Lejeune.....	67
Subtl.ai at the FinSBD-2 task: Document Structure Identification by Paying Attention Abhishek Arora, Aman Khullar, Sarath Chandra Pakala, Vishnu Ramesh and Manish Shrivastava.....	75
The FinSim 2020 Shared Task: Learning Semantic Representations for the Financial Domain Ismail El Maarouf, Youness Mansar, Virginie Moulleron and Dialekti Valsamou-Stanislawski.....	81

IITK at the FinSim Task: Hypernym Detection in Financial Domain via Context-Free and Contextualized Word Embeddings	
Vishal Keswani, Sakshi Singh and Ashutosh Modi.....	87
Anuj at the FinSim Task: Anuj@FINSIM–Learning Semantic Representation of Financial Domain with Investopedia	
Anuj Saini.....	93
ProsperAMnet at the FinSim Task: Detecting hypernyms of financial concepts via measuring the information stored in sparse word representations	
Gábor Berend, Norbert Kis-Szabó and Zsolt Szántó.....	98
FINSIM20 at the FinSim Task: Making Sense of Text in Financial Domain	
Vivek Anand, Yash Agrawal, Aarti Pol and Vasudeva Varma.....	104