FINANCIAL NAMED ENTITY RECOGNITION FOR TURKISH NEWS TEXTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

DUYGU DINÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JULY 2022

Approval of the thesis:

FINANCIAL NAMED ENTITY RECOGNITION FOR TURKISH NEWS TEXTS

submitted by **DUYGU DINÇ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Ali Hikmet Doğru
Supervisor, **Computer Engineering, METU** _____

Prof. Dr. Pınar Karagöz
Co-supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. İsmail Sengör Altıngövde
Computer Engineering, METU _____

Prof. Dr. Ali Hikmet Doğru
Computer Engineering, METU _____

Assoc. Prof. Dr. Mehmet Tan
Computer Engineering, TOBB University _____

Date: 26.07.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Duygu Dinç

Signature :

ABSTRACT

FINANCIAL NAMED ENTITY RECOGNITION FOR TURKISH NEWS TEXTS

Dinç, Duygu
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Ali Hikmet Doğru
Co-Supervisor: Prof. Dr. Pınar Karagöz

July 2022, 197 pages

Named Entity Recognition (NER) is a problem of information extraction where the objective is; in a given text, to detect and label named entities (NE) according to predetermined categories correctly. An NE may be a noun or a group of nouns which correspond to the name of a specific object, location or a concept in case of domain-specific applications. In the literature, person, organization, location names or date,time, money, percentage expressions are among highly studied, generic NEs. Besides, there are domain-specific studies with NEs that are related to specific domains like genetics, medicine, chemistry and finance. Solutions for NER problems may be useful in many downstream tasks in the Natural Language Processing domain such as Text Summarization, Question Answering and Sentiment Analysis. For Turkish, which has pretty complex morphological features, there are less number of studies in NER field compared to more widely used languages like English. In recent years, neural-network based methods performed better in NER tasks than classical rule-based or traditional machine learning techniques. In this thesis, most popular deep-learning based models were experimented using different Turkish datasets.

Moreover, as being one of the focuses of this thesis, from raw financial news texts, two newly annotated datasets were presented and used throughout the experiments. New datasets were annotated using both BIO schema and raw labels, inter-annotator agreements were measured and models were trained separately using both versions to observe the effect of annotation format on performance. Moreover, new NEs specific to finance were also presented. Lastly, experiments with a few selected deep-learning based language-specific BERT models for some languages in Ural-Altaic language group were conducted.

Keywords: deep learning, financial named entity recognition, named entity recognition, natural language processing, text mining

ÖZ

TÜRKÇE HABER METİNLERİNDE FİNANSAL VARLIK İSMİ TANIMA

Dinç, Duygu
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ali Hikmet Doğru
Ortak Tez Yöneticisi: Prof. Dr. Pınar Karagöz

Temmuz 2022 , 197 sayfa

Bir bilgi çıkarma problemi olan Varlık ismi Tanıma (VİT) probleminde amaç, verilen bir metin için, varlık isimlerini saptamak ve önceden belirlenmiş kategorilere göre doğru şekilde etiketlemektir. Bir varlık ismi (Vİ), özel bir objenin, lokasyonun veya alana özel uygulamalarda bir konsepti ifade eden isim veya isim gruplarından oluşabilir. Kişi, organizasyon, yer adları veya tarih, zaman, para ifadeleri literatürde en çok çalışılan varlık isimleri arasında yer almaktadır. Ayrıca, genetik, tıp, kimya ve finans gibi, belirli alanlardan varlık isimleriyle, alana özel çalışmalar mevcuttur. VİT problemlerinin çözümleri Doğal Dil İşleme alanındaki, Metin Özetleme, Soru Cevaplama ve Duygu Analizi gibi çalışmalarda da faydalı olabilir. Daha yaygın kullanılan İngilizce gibi dillerle kıyaslanırsa, oldukça karmaşık morfolojik özelliklere sahip Türkçe için daha az VİT çalışması bulunmaktadır. Son yıllarda VİT çalışmalarında, yapay sinir ağları tabanlı metodlar, klasik kural bazlı ve geleneksel makine öğrenmesi tekniklerine göre daha iyi performans göstermiştir. Bu tezde, en popüler derin öğrenme bazlı modeller ve farklı Türkçe verilerle deneyler gerçekleştirilmiştir. Ayrıca, tezin odaklarından birisi olarak, ham finansal haber metinlerinden iki yeni eti-

ketlenmiş veri kümesi sunulmuş ve deneylerde kullanılmıştır. Yeni veriler hem BIO şeması hem de ham etiketler kullanılarak etiketlenmiş, etiketleyiciler arası mutabakatlar ölçülmüş ve etiketleme şeklinin performansa üzerindeki etkilerini gözlemlemek için modeller her iki versiyonla eğitilmiştir. Ayrıca, finans alanına özgü yeni varlık isimleri de sunulmuştur. Son olarak, Ural-Altay dil grubundaki diller için eğitilmiş BERT modellerinden seçilen birkaçı ile deneyler yürütülmüştür.

Anahtar Kelimeler: derin öğrenme, finansal varlık ismi tanıma,varlık ismi tanıma, doğal dil işleme, metin madenciliği

To my beloved family

ACKNOWLEDGMENTS

I would like to present my deepest thanks to my supervisors Prof. Dr. Pınar Karagöz and Prof. Dr. Ali Hikmet Doğru who believed in me, motivated me with their positive energies, guided and leaded my MSc thesis work. It is a honour to be their students.

In addition, I want to express my gratitude to my thesis defence jury members Assoc. Prof. Dr. İsmail Sengör Altıngövde and Assoc. Prof. Dr. Mehmet Tan for evaluation and valuable feedbacks.

I would also like to express my special thanks to my beloved parents Melahat Dinç and Yakup Dinç, who raised me, always loved and supported me throughout all my journey. I would also thank to my precious sister Damla Dinç who always gave me the best wishes and supports.

Moreover, I would like to thank to my dear friends Gülşah Gürük, Gülin Duru, Özden Öztürk and Pelin Kılıç, for believing in me and for their support.

Furthermore, for all their support and kindness, I want to thank to Dr. Çağrı Kaya, Gökhan Özsarı and Fırat Çekinel from METU Computer Engineering Department.

I would also like to thank to two of my colleagues, Şaban Toros and Sümeyra Korkmaz.

Moreover, for providing the raw financial news texts (which formed a basis to our annotated FINTURK datasets) I would like to thank to Asst. Prof. Dr. Burcu Yılmaz from Gebze Technical University. In addition, I would also like to express my gratitude to the academicians and researchers (whose related studies were referred to in detail in the thesis) who took part in proposing the datasets that we named as TUR3 and EXTEND7 throughout our thesis. These datasets are valuable contributions for the literature since they were used as benchmark datasets in many studies including our thesis.

Lastly, I would like to thank to METU Computer Engineering faculty members who educated us in computer engineering field with patience and support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
CLM	Causal Language Modeling
CoNLL	Conference on Natural Language Learning
CNNs	Convolutional Neural Networks
CRFs	Conditional Random Fields
DeBERTa	Decoding enhanced BERT with disentangled attention
DARPA	Defense Advanced Research Projects Agency
FNE	Financial Named Entity
HMMs	Hidden Markov Models
IBAN	International Bank Account Number
IE	Information Extraction
KYC	Know Your Customer
LSTM	Long Short-Term Memory
MEMM	Maximum Entropy Markov Model
MLM	Masked Language Model
MRC	Machine Reading Comprehension
MUC	Message Understanding Conference
NE	Named Entity
NEE	Named Entity Extraction
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NSP	Next Sentence Prediction

POS	Part of Speech
QA	Question Answering
RNNs	Recurrent Neural Networks
RTD	Replaced Token Detection
SA	Sentiment Analysis
SBD	Sentence Boundary Detection
TLM	Translation Language Modeling

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

Named entity recognition (NER) is both a text mining and a natural language processing (NLP) problem where the aim is to extract prespecified named entities from the text given. NER studies can be domain specific or general depending on the dataset and named entity (NE) types.

An NE may be an entity which can be encountered with in all domains in general, such as "PERSON", "LOCATION" or "ORGANIZATION". On the other hand, there may be domain-specific NEs which are widely used in some domains such as medicine, genetics, finance, business, govermental subjects, agriculture, history, archeology, environment and so on.

Named entity recognition is indispensable for various high level semantic applications such as building knowledge graphs[8]. Other examples to downstream tasks which NER will be useful are; machine translation, question answering, sentiment analysis and text summarization.

Studies in the field of NER problem are concentrated mainly on western languages such as English, German, French etc. Of course there are also many valuable NER studies for widely spoken eastern languages such as Chinese. However, NER studies for Turkish are very limited where domain specific NER studies are much more limited than general NER studies. An important reason of this restrictiveness is the difficulty in finding Turkish text corpora.

Moreover, with the rise of deep learning based methods, an increase observed in the

performance in NLP tasks. This also applies for NER problem. In the studies for especially in English, very high F1-score results were reported. Since deep learning based methods gained popularity in more recent years, there are limited studies for Turkish NER problem with these methods compared to the classical methods based on other machine learning techniques.

Within the scope of this thesis work, our objectives were to propose brand new Turkish text datasets (originating from the same source) and apply current deep learning based algorithms to solve the NER problem that is domain specific.

We also aimed to conduct experiments with the datasets that were used in former NER studies for Turkish text in general domain and comment on the performances.

1.2 Proposed Methods and Models

In the scope of this thesis work, we aimed to evaluate performances of most popular eight deep-learning based methods using four different datasets, where two of them were newly created by processing raw text files in financial news domain and manually annotating them with newly-crafted domain-specific named entities. In addition, we presented a detailed annotation guideline that contains entity-wise comparisons and comments considering Seventh Message Understanding Conference (MUC). Moreover, deep learning based models were pre-trained for Turkish or multi languages and we indeed applied a fine-tuning process via training these models with our datasets at hand firstly and validating and testing afterwards.

Besides, different monolingual and multilingual models were trained (during fine-tuning) using both BIO annotated and NON-BIO annotated versions of same datasets and with these experiments, the result of annotation format on performance of a model is inferred.

Furthermore, k-fold cross-validation experiments were conducted for one of the datasets to observe the effect of distributional differences of NEs in datasets' train, dev and test splits which we used in the former experiments.

Moreover, using one of our newly presented datasets in financial domain we con-

ducted experiments with selected monolingual versions of pre-trained BERT models for different languages where a few of them are Ural-Altaic languages.

1.3 Contributions and Novelties

Our contributions are as follows:

- We proposed new annotated datasets in financial domain and these datasets which were annotated with different annotation formats can be experimented in further studies with different newer algorithms in the future.
- We experimented various popular deep learning based models to solve named entity recognition problem for Turkish texts with higher success compared to the current performance results reported in literature.
- We compared and commented on the results of different models on different annotated Turkish texts from both financial domain and general news domain.
- We evaluated the effect of annotation format (BIO/NON-BIO(raw)) on the performance of the model when exactly the same dataset was used other than the format.
- Considering results of k-fold cross validation experiments we observed and commented on the effect of the distribution of NEs in the different splits of datasets considering classical train-dev-test splitting process especially when the dataset is imbalanced
- We conducted experiments with language-specific BERT models for different languages where a few of them are included in Ural-Altaic language group.

1.4 The Outline of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, the results of literature review are mentioned. Information about the newly presented FINTURK dataset and it's annotation procedure with the details of named entities were explained together

with the background information about the models used in experiments,in Chapter 3. Experiments and results were presented and explained in detail in Chapter 4. Finally, in Chapter 5 Conclusions are expressed and draft plans about the future work were also mentioned.

CHAPTER 2

LITERATURE SURVEY

2.1 Named Entity Recognition

Named entity recognition is a task which can be considered as a subtask under other well-known task types in the NLP literature. For instance, NER is a subtask of sequential tagging. The task of applying tags to each token in a sentence consecutively is called sequential tagging.[9]

Another parent task of NER is the information extraction. Information extraction (IE) is the task of automatically extracting information of interest from unconstrained text creating a structured representation of this information. An IE task involves two main sub-tasks: the recognition of named entities involved in an event and the recognition of the relationships holding between named entities in that event [10]. Thus, named entity recognition (NER) or named entity extraction (NEE) can be considered as a sub field of information extraction. Named entity extraction task was introduced by Defense Advanced Research Projects Agency (DARPA)[11]. Studies of NER started in 1900's and studies for different languages took place in Message Understanding Conference (MUC) platform [12].

In MUC in 1998, a more elaborate definition of the task was presented: The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities. The task is to identify all instances of three types of expressions in each text in the test set and to subcategorize the expressions [13]. Similarly, in the scope of the CoNLL-2002 Shared Task [14], named entities are defined as phrases that contain the names of

persons, organizations, locations, times and quantities. Moreover, in CoNLL-2003 Shared Task [15] , named entities are defined as phrases that contain the names of persons, organizations and locations.

Successful solutions to NER tasks can facilitate studies in downstream tasks in many fields such as question answering, text summarization, recommendation systems and machine translation.

Named entity recognition can be divided into two sub fields as flat NER and nested NER. Nested NER problem, where a NE may also contain other NEs in itself, is beyond the scope of this thesis. Thus, from now on, as "NER", flat NER problem (in other words the traditional NER problem) will be discussed in the scope of this thesis.

There are different techniques throughout the literature to solve the NER problem. These techniques can be grouped into 4 categories:

- Classical rule-based techniques which benefit from lexicons, gazetteers or patterns extracted with pre-analyses of texts. Also, morphological features of the words in the language are sometimes used to strengthen the information resources. After their development using some specific domain data, these techniques are not very successful in inter-domain studies.

- Traditional machine learning techniques, using statistical techniques to build models from datasets.These methods are more dynamic than rule-based methods and models engineered with these methods can be adapted to different domains to some extent. Usage of CRFs in an NER task is a good example for this category.

- New generation deep learning based techniques became popular recently. With the power of the most popular subfield of machine learning, namely, deep neural networks and with help of some annotated and unannotated corpora, many NER applications achieved great successes.

- Hybrid techniques where both rule-based methods and some machine learning methods are applied together.

2.2 Named Entiy Recognition for Turkish

NER is an extensively studied problem in especially for some Western languages. NER task is known to be a solved problem for English with state-of-the-art performance above 90% [16].

Throughout the past decades, named entity recognition problem was not very frequently studied for Turkish compared to English or many other prevalent languages. One reason may be the fact that the techniques applied for the solution of this problem for English can be applied similarly to the languages in the same language family but not for languages that are not very close to these European languages genetically or typologically. Another strong reason is challenges in applying conventional methods that depend on lexical and relatively complex morphological characteristics of Turkish. Turkish has very productive inflectional and derivational processes. Many local and non-local syntactic structures are represented as morphemes which at the end produces Turkish words with complex morphological structures and that leads to high productivity and cause data sparseness problems[17]. Yet, as the technological developments increase rapidly, more effective machine learning and especially deep learning-based methods were studied and provided spectacular successes in NLP and its sub domains such as NER.

2.2.1 Place of Turkish Among Languages

Turkish is a morphologically complex and agglutinative language. Due to its characteristic features, unlike some Western languages such as English or German, words are reproduced via adding derivational and inflectional affixes to the root form or the body of a word depending on the situation. Derivational affixes are the ones which makes the word that it is added, gain a new meaning with this affix. For example, adding the derivational affix "-lük" to the word "göz" (eye), the word "gözlük"(eyeglass) is constructed. After adding the affix "-çü", the word "gözlükçü" (optician) is obtained. On the other hand, inflectional affixes do not cause the word gain a totally new meaning but they provide possessive, case, plurality, tense and person relations to the context. For example, if the affix "-üm" is added to the word

"göz" (eye) the word becomes "gözüm" (my eye) where not a new word with totally a different meaning is constructed but the existing word gained a possessive meaning. Similarly, adding "-lar" affix to the word "kitap" (book) results with the word "kitaplar" (books) where the affix does not completely change the meaning of the beginning word "kitap" but added a plural meaning to it. Because of this agglutinative structure, many other words can be obtained via adding derivational and inflectional affixes to existing words.

Richness is another property of Turkish. As Tosun[18] mentioned in their enlightening paper, three important Turkish linguists and scientists agree that there are some criteria that prove the richness of a language:

- The richness in vocabulary,
- The richness in concepts,
- Having the most ancient written works and
- Having been spoken in large geographical area.

As Tosun also stated in the scope of their study, according to Turkish Language Association, in the 11^{th} edition of the Turkish dictionary, there are 122,423 vocabulary items consisting of words, terms, idioms, affixes and meanings, a total of 92,292 words per article and within the article, 34,672 sample sentences selected from Turkish literature, and a dictionary text consisting of 1,454,903 words. For the second criteria, Tosun [18] gave the example of richness in the color names and detailed description feature in kinship relations. As the author stated, there are words expressing the kinship relations as "kayınbirader" (brother-in-law) , "bacanak" (husband of sister-in-law), "enişte" (brother-in-law or uncle-in-law) , "elti" (sister-in-law), "yenge" (aunt-in-law), "görümce" (sister-in-law, husband's sister) and "baldız" (sister-in-law, wife's sister), where these concepts are met with a single indicator according to gender in French, German and English. These examples of richness in the vocabulary can be extended with "amca" (uncle, father's brother), "dayı" (uncle, mother's brother), "hala" (aunt, father's sister), "teyze" (aunt, mother's sister) where these 4 different words corresponds 2 different words in English as seen. Idioms are

other examples of the richness in concepts. According to the Turkish Language Association [19], there are 2,396 proverbs and 11,209 idioms in the proverbs and idioms database of the institution. Moreover, in the paper [18], Göktürk and Orhun inscriptions were provided as evidences to the third criteria. Lastly, for the last criteria, it was stated that Türkiye, Turkmenistan,Azerbaijan,Uzbekistan, Kyrgyzstan, Kazakhstan, Turkish Republic of Northern Cyprus , East Turkestan, Chuvasya, Khakas, Tatarstan, Tuva, Dagestan,Bashkortostan, Mountainous Altai, Gagauzia, Kabardino-Balkaria, Karachay-Cherkessia, Afghanistan, Syria, Iraq, Iran, Tajikistan, Ukraine, Mongolia, Macedonia, Bulgaria, Yugoslavia, Greece, Albania, Romania, Germany and other European countries, America and Australia were counted as the geographical areas that Turkish is spoken.

Besides the richness in vocabulary criteria mentioned above, Aksan [20] claimed that the power of expression is as important as the breadth of the vocabulary in the richness of a language, or even more than that. Author stated that many concepts in the Turkish public language and dialects of the Turkish language have been found by making use of the fertility of the language. For this case, author provided some examples from disease names that Turkish people named using the distinctive natures of diseases such as "kuşpalazı" (a compound word including "bird" (kuş) and "poult" (palaz) words) and "küf kuyruğu" (tail made of mold) for diphtheria.

Aksan also claimed that there is a sterility in words of some Indo-European tribes, Greeks and Latins, related to colours which causes some linguists to doubt whether these tribes are colorblind or not. However, it was also mentioned that Turkish shows a great abundance in colour names, starting with Orkhon Inscriptions. According to the author, in Indo-European languages, colour names such as green, wax, gray and yellow, often seemed to be derived from a single root. All these adjectives meaning colour are connected to the common root meaning "to shine, to give light" and are seen as related to each other. On the other hand, as seen in the oldest documents of Turkish, these colour names were derived from different roots and different ways of expression. For example, "yaşıl" ("yeşil" in modern Turkish) for green and "kök" ("mavi" in modern Turkish) for blue in Kültigin Inscription, "sarig" ("sarı" in modern Turkish) for yellow in Tonyukuk Inscription.

Because of the sparsity and richness in the language, preparing lexical or morphological rule tables or applying rule-based systems for NER is not very easy for Turkish since there are too many possible combinations of words and affixes. However, more dynamic and learning based methods can be successful in such rich and complex languages.

After the success of deep learning based models such as BERT model [2] on English NER tasks, researchers began to apply multilingual BERT for texts in their languages and received generally good results which are maybe not as good as state-of-the-arts in those languages. Since multilingual BERT was trained with 104 different languages its success on different languages is up to a certain extent. However, for many languages including Turkish [21], Finnish [22],Estonian [23], Hungarian [24], Korean [25], Greek [26], German [27], Arabic [28], French [29], there are language specific BERT-based models which were developed via training BERT with larger corpora from those languages.

Since named entity recognition is a natural language processing task and since today there are some widely accepted linguistic theories stating there are language families and some of the languages are classified as close relatives of each other, the comparative performance of deep learning based models on NER tasks for various languages may be strongly dependent on the structure of these languages. In this respect, investigating the relation of Turkish with other languages, in other words, its place among other world's languages is seemed to be important.

In terms of formation, languages are categorized as monosyllabic languages, flexiomal languages and agglutinative languages [30] (as cited in [31]) where Turkish is under the agglutinative category. Moreover, author pointed out another taxonomy of languages based on "family" perspective. According to this point of view, "a language family is the name given to a group of languages, formed by languages separated from the main languages by the way of development and change". As stated in the paper, there are Indo-European, Semitic, Bantu, Chinese-Tibetian language families where Turkish belongs to none of these families but to Ural-Altaic language group which is not considered as a family due to the fact that the closeness seen in Indo-European languages cannot be observed in these languages. Author mentioned

that the closeness between this group of languages come from the structure unity. Moreover, Ural-Altaic language group is divided into Ural and Altai branches where Finnish, Hungarian, Uygur languages belong to Ural branch and Turkish, Mogol languages belong to Altai branch. Another important point that the author put forward is that there are some linguists who decline this theory for Ural-Altaic languages and there are some views who state that the relation between Altaic languages come from exchanges of words and cultures.

Another comprehensive paper about the place of Turkish among other languages was written by Ercilasun [32] where the author mentioned about two main classifications of the world languages such as “genetic” and “typological” classifications. Turkish belongs to Altay languages along with the Mongolian and Manchu-Tungus languages from the viewpoint of genetic classification. It was stated that Altaists include Korean and Japanese to Altay language family and they argue that ancestor of these languages is common. Besides, it was also stated that some Turcologists refuse common ancestor theory between Altay languages and assess that the similarities come from interaction between these languages. Considering typological classification the main criteria is word-building criteria and linguists classified languages as “Isolating / analytic languages”, “Agglutinating languages” and “Inflectional / fusional languages”. As the author explained, since in Turkish, words are formulated by adding the morphemes with each other and since the relations between the words also depend on endings and lastly since there is no ablaut (namely, words "kış" (winter) and "kuş" (bird) are totally different words), Turkish belongs to agglutinating languages category. According to the author, *“Agglutinating or inflectional/fusional characteristics don’t denote a kinship between languages. There are a lot of languages which are unrelated with each other and are classified as agglutinating languages.”*. However, author noted that languages in same family expose same characteristics. For example, in Turkish sequence of words in the sentence follows “subject-object-verb” order such as in other Altay languages.

To sum up some different viewpoints about the resemblence of Turkish to other languges, it can be said that similarities between Turkish and Altaic languages may be higher than the resemblences between Turkish and other languages. This degree of similarity is followed by the similarity between Turkish and Uralic languages. Since

Ural-Altaic language group is not accepted as a family, we may not expect very strong similarity between Turkish and these languages in the group but it is worth to compare the results of experiments on NER tasks in these languages to the results of NER experiments for Turkish. To say the least, at the first sight, finding similar success levels in NER task for these languages in the same group is not so unexpected.

2.2.2 Studies of Named Entity Recognition for Turkish

In 1999, Cucerzan and Yarowsky published a paper [33] where they handled the NER problem independent from the languages, using morphological and contextual information. In their paper, they mentioned the obstacles in obtaining annotated datasets and proposed a *"bootstrapping algorithm based on iterative learning and re-estimation of contextual and morphological patterns captured in hierarchically smoothed trie models"*. They conducted the experiments for Romanian, English, Greek, Hindi and Turkish. With full bootstrapping method they applied, beginning with some seed names for entity types, contextual patterns and morphological characteristics were learnt and they obtained an F-score of 53.04 for Turkish where the lowest F1-score was 41.70 which was obtained for Hindi and the highest F1-score was 70.47 which was obtained for Romanian.

One important study of NLP with Turkish texts is by Tür et al. [11] in 2003 where statistical language processing methods were applied during the extraction of information. Authors studied on three tasks: Sentence segmentation, topic segmentation and name tagging. Name tagging task can be explained as detecting and labelling only the "name" entities-person, organization and location entities- in other words, enamex entities in a classical named entity recognition task. Moreover, as they stated in their paper, for name tagging task, they modeled n-gram language models embedded in Hidden Markov Models. Name models were constructed with different information sources and lexical, contextual, morphological and name tag models were proposed where at the end with combination of this models they obtained 91.56% F-score which was a substantial improvement considering the success in information extraction field for Turkish back then. In lexical model, one vital thing that increased accuracy was detecting boundaries of names where it made the probability of label-

ing named entities which consists of more than one word, higher. Furhermore, in the contextual model, to be able to infer from the context, they made use of the lexical model. Simply, they tagged the word to be tagged as "unknown" and and using the context, surrounding words, estimated the probability of the word being in classes of person, organization, location or else named entities. Another model they presented for name tagging was the morphological model. For this model, they benefited from morphological analyses of the words. For a word, they firstly extracted all morphological features of that word, formulated that as an output sequence and calculated the probability of the word being a name. However, since this model cannot distinguish the type of a name entity, they tagged the words as person and then combined this model with the other models they proposed. Furthermore, they presented the tag model where there are person, organization, location and else tags and boundary flag types (yes, no and mid). As they mentioned, to be able to use the tag model, they needed the posterior probabilities obtained from any combination of lexical, morphological and contextual models. This posterior probability would be needed to get state observation likelihood using Bayes rule and Viterbi decoding. Then, transition probabilities were obtained. In experiments with Turkish texts they reported F-measures between 86.01% and 91.56% for various models and their combinations.

In 2008, Bayraktar and Taşkaya-Temizel [34] published a paper on the person name extraction from Turkish financial news text using local grammar-based approach. They tried to find reporting verbs in Turkish texts to reach proper nouns that are very close to these words in Turkish financial news texts. After finding out reporting verbs, they categorized them into 3 different formats and conducted analyses to find different patterns which include reporting verbs and person names in different positions relative to each other. They also utilized from the capitalization rule which finds all capitalized words in the text as a naive proper noun extraction technique. According to the reported results, local grammar-based approach with capitalization rule performed much better (with 81.97% F-measure) than the approach using only capitalization rule (F-measure as 12.48%).

In 2009, Küçük and Yazıcı published a study of NER for Turkish texts from news domain where they developed a rule-based system [35].They utilized from lexical resources such as dictionary of person names and a few lists of well-known locations,

people and organizations. They also made use of pattern bases for location and organization names and temporal and numeric expressions. They obtained 78.7% F-score after the evaluation of overall test samples. For the evaluation on news text, authors mentioned that the rule-based system had difficulties in recognizing person names just using the dictionary and by looking their examples in the study it can be said that many person names in Turkish are homophones of words that are not proper names. Similary, system had difficulties in recognizing organization and location entities and the homophonic words may be one reason for that as well. They also mentioned that extraction of compound organization names are not foolproof. Authors compared their study with [11] and concluded that *"the statistical system performed deeper language processing compared to the rule-based system"*. Moreover, authors evaluated this rule-based system on child stories and historical texts as well and observed F-measures of 69.3% and 55.3% respectively. They reported that one of the disadvantages that degrades the performance for child stories is the existence of foreign person names in the texts which does not exist in lexical resources. Similar problem was also reported for historical texts, due to the fact that the historical organizations and people do not exist in lexicons, the performance in extracting them decreased. Furthermore, as a fourth different data set, they evaluated their rule-based system on news video transcripts and obtained 75.1% as F-measure which was significantly higher than in case of child stories and historical texts and expected since video subjects are news. As the authors mentioned in the conclusion, rule-based systems do not require the usage of training data, however on the other side, they are not dynamic as machine learning and deep learning based techniques since language is not a static thing and keeps growing and changing by time. Another disadvantage of rule-based systems is the hardness to apply them on different fields since the lexicons and pattern bases are constructed for the original domain.

After rule-based NE recognizer, Küçük and Yazıcı presented a hybrid NE recognizer [16] for Turkish which they built upon the rule-based NE recognizer [35] and applied it for various types of Turkish texts such as financial news texts, child stories and historical texts. They equipped the rule-based NE recognizer with a rote learning component and obtained a hybrid NE recognizer. The F-measures were 83.81%, 59.05%, 73.97% and 60.96% for news text, financial news text, child stories and

historical data sets respectively. This first hybrid NE recognizer for Turkish texts showed higher performance compared to the rule-based system proposed before [35], since it has an additional information resource, rote learner which used statistical analyses to extract entities.

Moreover, Yeniterzi's paper [17] showed the gain from morphological analyses in NER for Turkish. An interesting side of the study is that besides the word-level representation, roots and morphological features were represented as separate tokens too. Author stated that this approach provided a 7.6% relative improvement over the baseline. Also, it was mentioned that usage of Conditional Random Fields (CRFs) was preferred since they are advantageous compared to HMMs. In the paper, two different models were proposed, word-level and morpheme-level models. Moreover, there are root, part-of-speech (POS) tag, proper noun and case features that are used in two different models. In overall, morpheme-level model showed slightly more success than word-level model and all of the features improved the system significantly.

In 2011 Özkaya and Diri published a study of a rule-based NER with CRFs [12]. As dataset, they used e-mails (academic, institutional and personal e-mails) in Turkish which were written using informal language and aimed to extract and tag location, organization and proper names. For every token in the e-mail dataset, some features were extracted such as length of the word, cardinality, part of speech tag, being the first word of the sentence and also some features from the e-mail's fields such as "subject" were extracted. Words were recorded as a list to be used as an information resource. After the feature extraction, tag was determined with CRF method and using rule definitions. Besides, authors also used some external lexicons which contain list of abbreviations, titles and proper names. Authors observed that the most successful NER was for institutional e-mails and for the person named entity. This result is not very surprising considering that institutional e-mails may be more formal than an average e-mail from any domain. And this study may be thought as an important example for the early NER studies with less formal Turkish texts.

Furthermore, in [36] Tatar and Çiçekli proposed an automatic rule learning method that benefited from orthographical, contextual, lexical and morphological features. They experimented on TurkIE dataset which was a corpus of articles collected from

different Turkish newspapers. They observed an average F1-score of 91.08% on a dataset that includes person, location, organization names besides date and time as named entities.

Another study with CRFs as the statistical model was by Şeker and Eryiğit [37]. They benefited from the morphological features, lexical features and gazetteers to recognize person, organization and location entities in general news texts as the dataset which is also the dataset used by Tür et al. in [11] withal they achieved 95% in MUC and 92% in CoNNL metric.

Another Turkish NER study using is by Eken and Tantug[38] who published a paper on recognizing named entities in Turkish tweets. In their paper, they refer to the hardness of processing informal texts such as e-mails, microblog texts and social media texts since they may be ungrammatical and may include spelling mistakes unlike the formal texts. Authors used news data to train their base model and to train their second model they used tweets which are short texts lack of context, containing spelling errors, slangs and repeating characters. They used CRF to build their model. They used first and last 4 characters of the word, capitalization and apostrophe information as morphological features and they also benefited from gazetteers where at the end, they observed 64% F1-score.

In 2017, Şeker and Eryiğit enriched their former study, proposed a newer model based on CRFs again and constructed brand-new datasets in [39] to become able to use less formal user generated content data from Web 2.0 and increased the types of named entities to 7 via adding "Date", "Time", "Percentage" and "Money" entities to "Person", "Organization" and "Location" entities that are available in the former study[37].They observed an F1-score of 92% with the Turkish news dataset and approximately 65% F1 score for new datasets which were obtained from web and annotated. These results exhibit that there is a big gap between the success level of NER solutions for Turkish formal and informal texts.

In [40], Çelikkaya et al. proposed three newly annotated Turkish NER datasets from Twitter, a speech-to-text interface and a hadware forum. These datasets contain NEs: Person, organization, location, date, time, money and percentage. With their three different feature models, they conducted experiments on these datasets and reported

the results which showed the difficulty of NER using informal texts.

Classical rule-based techniques and traditional machine learning techniques solve the NER problem up to some extent (using the features that are built using some rules or based on some patterns and gazetteers) but for mainly the formal data such as news texts where there are approximately no missing words/characters, slangs. Since these formal texts are well structured, solutions crafted experimenting on these types of formal datasets do not apply well to the solecistic informal texts such as tweets or texts from social media domain which may contain incomplete sentences, errors in punctuation and slangs. Hence, a more dynamic methods, such as neural network based methods, are needed to solve NER problem with a high success rate when it comes especially to informal texts.

In the recent years, neural network based models showed higher performances in sequence tagging tasks such as NER. In 2014, Demir et al. [41] handled the NER problem with a neural network based semi-supervised learning approach. In the unsupervised stage, to taught the continuous word representations, researchers used the skip-gram architecture and trained the proposed model with large amount of unlabeled data collected from several Turkish news sites and publicly available Chezch data crawled from news sites. In the supervised stage, authors experimented on a Czech dataset CNEC 1.1 prepared formerly [42] besides the Turkish dataset which was firstly used in [11]. With their final system F1 scores of 91.85% and 75.61% were observed for Turkish and Czech respectively. Since their proposed system does not take advantage of the language dependent features, they assessed that that system can be useful for many different morphologically rich languages.

In 2016, using a deep bidirectional LSTM, Kuru et al.[43] proposed a character level NER system where the sentence is represented as a sequence of characters instead of sequence of words and tags for each character is estimated with some probability. Afterwards, with Viterbi decoder they transformed these probabilites to word labels. Their system achieved a phrase level F1-score as 91.30 for Turkish.

In 2019, Çekinel et al. proposed a model where they combined CRF with BERT-Base language model, besides they investigated the effect of new features on model accuracy. They used Turkish datasets which were used formerly in[11] and [39]

where datasets contain 3 and 7 generic NEs, respectively. For the first dataset[11], in overall, they obtained 90.15% F1-score. They also used the social media dataset that was firstly proposed in [39] as an additional test dataset and obtained. For the second dataset (from [39]), since they could obtained just the train set of the original dataset, they conducted 10-fold-cross-validation experiments with this part of the original dataset and obtained an F1-score of 91.74% in average. Moreover, authors concluded that adding new features increased the model's performance. In addition, their proposed system received relatively high performance results even if they did not use any lexicons.

In [44]Güngör et al. investigated the effect of morphology in NER and proposed a sequential tagger,including a bidirectional LSTM layer followed by a conditional random field layer, for NER task in morphologically rich languages such as Turkish. In addition, they conducted experiments with Czech, Hungarian, Finnish and Spanish as well. They emphasized the importance of understanding the morphological characteristics for morphologically rich languages via stating that they *"retain information in the morphology of the surface form of words which is present in the syntax and word n-grams in other languages"*. Because of this reason, they thought that morphological tags holds semantic and syntactic information for the language and hence useful in the solution of NER. During the training and evaluation phases of their model, the Turkish dataset that was used in [11] was used. Their system received 92.93% as F1-score for Turkish NER.

More recently, a very popular transformer based architecture, BERT[2] was proposed -which will be mentioned in the following chapter in detail-and in our paper [45] we investigated the performance of BERT on Turkish news texts and using the multilingual BERT, obtained overall F1-scores of 80.63% and 82.31% for datasets used in [11] and [39], respectively. Moreover, BERT model was trained for Turkish[21] as well.

Moreover, Akkaya and Can[46] experimented using informal/noisy user-generated data. They extended a BiLSTM-CRF model by adding an additional CRF and proposed a transfer learning model. Authors mentioned the scarcity of annotated noisy data in Turkish. As authors explained, since the CRF layer included two CRFs where

one was trained on large formal texts and the other one was trained on noisy texts, the dependencies learnt from larger formal texts were transferred to noisy text side. Thus the performance of the model considering the rarely-seen entity types (such as TIMEX and PERCENTAGE entities) increased. They obtained F1-scores of 67.39% and 45.30% for Turkish and English noisy data, respectively.

Furthermore, Aras et al. [47] compared the success of Bidirectional Long Short-Term Memory (BiLSTM) and transformer-based networks on NER task and found out that transformer-based networks outperform BiLSTM networks.Moreover, authors proposed a transformer-based network with a CRF layer (BERTurk-CRF) and they received an F1 score as 95.95% on the dataset that was used firstly by Tür et al.[11].

2.3 Selection of Named Entity Studies for Other Ural-Altaic Languages

As stated in section 2.2.1, Ural-Altaic languages is a language group rather than a language family. In despite of their distant proximities, some resemblence is expected between these languages since they are in the same group. This leads us an idea that the performance with deep learning based models (such as BERT, ELECTRA etc.) may also show similar performances for these languages in the same group.

In 2021 Ács et al. [8] conducted series of experiments using various multilingual and language-specific BERT models for Uralic languages including Estonian, Finnish, Hungarian, Erzya, Moksha,Karelian, Livvi, Komi Permyak, Komi Zyrian, Northern Sámi, and Skolt Sámi and provided the performance metrics for three NLP tasks including NER. They found out that the the performance difference between native BERT models and XLM-RoBERTa model is not statistically significant.Another key finding is that for NER, comparing with multilingual models, using native models with closely related languages does not provide substantial improvement considering performance. For NER, they used 2000 training, 200 validation and 200 test sentences in maximum, when available for different languages. Again, considering NER task evaluation, for Uralic languages which are using Latin alphabet, namely Hungarian, Estonian, Finnish and North Sami, multilingual BERT is the best performed model

for Estonian and North Sami and for Hungarian and Finnish, their language-specific BERT models performed best which is followed by multilingual BERT. Authors also compared NER results on languages that use the Cyrillic alphabet, namely Erzya , Moksha, Komi Permyak and Komi Zyrian. These languages do not have language-specific BERT models. For the first 3 of these languages, RuBERT (Russian specific BERT) performed the best in NER task and there is a small difference in performance of RuBERT and multilingual BERT. For Komi Zyrian, multilingual BERT performed the best. The success of Russian specific BERT is not highly surprising since these languages may have been affected from Russian because of geographic proximity. Lastly, since English is not close to Finnish, Hungarian, Estonian and Nort Sami languages considering their closeness degree between each other, it is expected to see lower performances with EngBERT model compared to other language models. However, for Finnish, EstBERT and HuBERT performed worse than EngBERT; for Hungarian, EstBERT and FinBERT performed worse than EngBERT and for Estonian, the performance of EngBERT is close to the performances of FinBERT and HuBERT.

Finnish is under the Ural language branch of Ural-Altaic language group. There are many NER studies with Finnish texts. In 2019, Virtanen et al. [22] presented a study for Finnish which is a "lower-resourced language" (such as Turkish) compared to English. They conducted experiments for both multilingual BERT and also trained a new Finnish BERT model as well. Authors used a dataset containing person, organization, location, product, event and date NE's. For NER task, new model (FinBERT cased) outperformed formerly presented methods with F1-scores of 92.40 and 81.47 for in-domain and out of domain test sets respectively. They also concluded that multilingual BERT model fails to outperform former state-of-the-art methods and so, is not successful in deep transfer learning for lower resourced languages.

Estonian is another language under Ural language branch. In 2020, by Tanvir et al. [23] it was also mentioned that Estonian is a less-resourced language and EstBERT model was obtained via pretraining BERT with a large Estonian corpus. Researchers pretrained EstBERT for multiple NLP tasks which include NER. In their paper, they also compared the performance of EstBERT on NER task with performances of; another Estonian specific BERT model (WikiBERT) which was trained using a smaller

dataset, multilingual BERT model and XLM-RoBERTa model. For the NER task they concluded that EstBERT is not the best performed model in overall but it outperformed multilingual BERT and XLM-RoBERTa. The best performed model was WikiBERT model but there is just a small difference in F1-score between both Estonian specific NER models.

Hungarian is another member of the Ural languge branch.In 2021, Nemeskey [24] introduced HuBERT which is a language-specific BERT model family for Hungarian. The HuBERT model which has the same name with the model family, was evaluated using masked language modeling and outperformed the multilingual BERT. It was also reported that the proposed model achieved state-of-the-art results for NER task with F1-score of 97.62%.

Mongolian is under the Altaic branch of Ural-Altaic Language group and is among languages which exposit being low resourced. In 2020,Cheng et al. [48] presented a corpus for Mongolian NER. They mentioned the scarcity of NE types specific to tourism domain other than the classical ones such as person, location, organization. Their corpus includes 16,000 sentences and 18 different NEs. They obtained an NER model by pre-training BERT Base model using a Mongolian corpus of 10 GB and then trained their NER model with Mongolian Tourism NER corpus. An overall 82.09% F1-score was observed with an increase of 3.54 points compared to traditional CRF NER method. Their annotation also contains coarse-grained 5 entity classes which are hierarchically over these fine-grained 18 classes. Researchers formed their dataset using news, essays, scenic spot intros, travel notes and other materials. They adopted BIOES schema and also mentioned the inter-annotator agreement as 0.85, measured using Cohen's Kappa.

Uzbek is another Altaic language. In 2021, Mansurov et al. [49] proposed a language specific BERT model for Uzbek (UzBERT). Considering the masked language model accuracy, their model performed much more better than the multilingual BERT for Uzbek texts. Authors mentioned that in general, monolingual models perform better than their multilingual counterparts and also they include smaller vocabularies and less parameters than multilingual models and so less computing power is enough to achieve fine tuning. In the paper, it was also assessed that since Uzbek is not a

high-resource language, there are no publicly available datasets for downstream tasks including NER and for that reason authors could not evaluate the performance of UzBERT model on NER.

2.4 Domain-Specific Named Entity Recognition Studies in Financial and Related Domains

While the volume of data produced in finance and economics is constantly increasing, it is technologically possible and valuable to use this precious data to analyze future trends and reality of current expectations for market parameters. Obtaining labeled financial datasets is much more costly and time-consuming since annotation of this kind of datasets require financial knowledge and background. Thus, it is important to take advantage of deep learning based models and transfer learning with the limited supervised data available.

In the literature, there are many NER studies with data from financial domain, however most of these studies are for languages other than Turkish. One of the earliest NER studies with financial data was published in 1970 by Farmakiotou et al. [10] which was a rule based study tested on Greek financial news corpus. As authors stated in their article, their NER system was based on hand crafted lexical resources such as grammars and gazetteers. There were 3 categories of NE's, persons, organizations and locations. They achieved 0.869, 0.824, 0.816 as F-scores for organization, location and person named entities, respectively. An important reason for the less success in person NE was due to the 33% of non-Greek person names in the evaluation corpus. Furthermore, as they mention in their paper, more than 30% of the mistakes in the organization NER proceeded from spelling and preprocessing errors where preprocessing contained tokenization, sentence splitting, POS tagging, stemming and matching against lists of known names.

Another NER study for Greek financial texts was published in the year 2000 [50]. In the study, a named entity recognizer for Greek was offered and a corpus of Greek texts of 12M words (from financial newspapers, financial portals and financial magazines) was utilized. Named entities in the study were person, organization, location, date,

time, percent and money. Proposed system consisted of an automated pipeline for text processing via pattern matching where and obtained F-score was 0.83 which was not low considering that deep learning based methods were not used.

In 2014, Wang et al. [51] conducted a study with Chinese financial news dataset that was annotated by authors themselves. As they stated in their paper, until when, many approaches faced with difficulties in extracting financial named entities (FNE) especially for the abbreviation of FNE's. Their dataset included stock names and some specific financial terms and some of the NE's (approximately 15% of the FNE's) were in abbreviated form. Firstly they achieved recognition of full form named entities with a model based on CRFs and then by making use of the recognized full FNE's with mutual information, information entropy and word similarity measurements they recognized the abbreviated entities with success. They achieved 0.9102 and 0.9277 for precision and recall metrics respectively.

Another NER paper in financial domain was published in 2016 by Emekligil et al.[52] where noisy unstructured customer order texts in Turkish (fax-image formatted documents converted to texts) were used. Besides, banking sector related named entities such as client name, organization name, bank account number, IBAN, amount, organization, currency and explanation were extracted via a CRF based NER technique. As they stated, banks receive a lot of customer transaction orders via fax channel and because of the difference in formats and the image resolution difficulties, after fax orders turned into text files with optical character recognition technology, obtained dataset is highly noisy. With presented system, they aimed to make use of NER to automatically extract entities which are required to complete customers' financial orders from customers' image type files and minimize the human effort necessary to read and input the customer order details from fax document to the core banking system. Actually, considering the banking sector in Turkey, successful complete application of this technique in operations centers of commercial banks will reduce the personnel expenses and increases profitability. As authors mention, they used linear chain CRFs that solve the feature restriction problems that exist in Hidden Markov Model (HMM) and label bias problem in Maximum Entropy Markov Model (MEMM) by combining good sides of both models. Moreover, as they explained in detail in their paper; they selected features by looking at most frequent n-grams, skip-grams person names,

punctuation, word shape. Authors received an average F-score of 72.77% on test set with their final model which was significantly lower than the models trained with the news data, as they express clearly in their paper. They also expected the result to be worse than the news domain since their dataset is noisy which was constructed via processing unstructured and poor resolution fax documents with OCR.This is an important study for both Turkish NER and Turkish finance-banking domain. After this study's implementation to the bank's core banking system, 20% of the manual workforce was saved and overall cycle time of target processes such as EFT was drastically decreased as they explained with examples in the paper. In the future, if there will be a standard pre-designed customer order format available in banking sector, these kinds of NER studies will give much more better results.

Getting closer to present, deep learning-based models became much more popular and applicable since the former problems in the limitation of resources such as CPU, GPU and memory has been decreased with the substantial advances in the technology. Another resource problem was the availability of large annotated corpora which is still an important limitation when studying with deep learning based procedures. It is really challenging to find datasets in a field of interest. During the studies for this theis work, same dataset problem arised and thanks to my advisor, and another academician who provided us the raw-unannotated news texts, we at least obtained raw financial news texts in Turkish. As described in the dataset section, the raw dataset was manually annotated by us for a long time. Putting these difficulties aside, with the ease of the restrictions in memory and processing, computer science world is now making use of powerful deep learning methods in various sub fields. While the success of deep learning in natural language processing (NLP) problems, NER tasks have been started to be solved with new deep learning based or hybrid (classical lexicon-based or similar techniques along with deep learning techniques like an ensemble) solutions.

In 2019, Francis et al. [53] showed that transfer learning can be used to solve the problem of scarce annotated data in a target domain. They demonstrated that *"NER models trained on labeled data from a source domain can be used as base models and then be fine-tuned with few labeled data for recognition of different named entity classes in a target domain."* and they used BERT as a languge model to improve their

biomedical NER model. In this study, they extracted NEs from documents in both financial and biomedical domains. In financial case, they used 3000 business documents such as invoices, business forms and emails. They used 60% of the dataset as training set, 25% as test set and 15 % as development set. They aimed to extract invoice sender name, invoice number, invoice date, International Bank Account Number (IBAN), total inclusive amount, total exclusive amount, company name, company address etc. as named entities. Since the solution of an NER problem is very beneficial for real-life problems, they mentioned that *"the overall goal of this use case is to research and test prototypes for a self-learning SaaS platform for simplification of data-intensive customer interactions in industries such as energy, telecommunications, banking or insurance"*. To recognize financial named entities authors used two networks, first was a filter network to detect the segments of texts whose probability of containing named entities of interest is more than others. These more probable text lines are then passed to the second network which is a sequence labelling (NER) network containing input layer, 2 bidirectional LSTM layers with batch normalization applied in between and a softmax CRF layer on the top. As the output of this network, named entities' classes and their span were predicted. As the source domain labels they chose IBAN, invoice date, total inclusive amount and invoice number and as the target domain labels they chose invoice date and expiration date. With the transfer learning, even if they trained with lesser labeled target data they observed considerably higher F1-scores than without the transfer learning case.

A domain-specific template mining study was by Sagheer et al. [54] in 2019 which also included an NER study in it as a step. Their domain is oil-production and the language is Arabic but their named entities are considered to be related to finance through production. There are NEs such as: Organization, underground wealth, numerical statue, number, level, time, unit, country and political person. They used Arabic text corpus from Arabic BBC, RT Arabic and Arabic CNN newspapers. First of all, researchers analyzed text morphologically and determined "verb" NE with this analysis. On the other hand, they conducted a semantic analysis by using a lexicon and determined many of the NEs with the semantic analysis. After NER, they used a rule-based system providing recognized named entities for mining of text templates. They set up 9 rules to mine different templates from paragraphs. For example, one of

the rules is activated when the paragraph is telling about quantity of organization's oil production and another rule is activated when the paragraph is about the levels of oil production. With the successful accomplishment of template mining task is valuable for down-stream tasks such as text summarization or question answering as authors also stated in their publication.

Even if it was not trained specifically for an NER task, since it was a financial BERT-based model, one of the worth mentioning studies here is the Liu et al.'s [55] study in 2020 where authors presented FinBERT (BERT for Financial Text Mining) model which is a domain specific language model pre-trained on both large-scale financial corpus and general corpus to make the model capture language knowledge and semantic information. As they mentioned in their paper, there are 6 pre-training tasks in FinBERT and their model showed better performance compared to state-of-the-art models at that time. They proposed both large and base versions of FinBERT. Authors did experiments with their newly proposed model on financial benchmark datasets and concluded that FinBERT was better in financial Sentence Boundary Detection, financial Sentiment Analysis, financial Question Answering tasks compared to state-of-the-art models. As pre-training data, they used English corpus from BooksCorpus (Zhu et al., 2015) and English Wikipedia as general domain data which was 13 GB. Besides, for the financial data, they crawled on Internet on websites in financial domain. Totally they had 48 GB financial text corpora from five resources including financial articles, financial news texts, question-answer pairs about finance from a website. Researchers trained FinBERT model from scratch on their 61 GB unsupervised corpus and then fine-tuned it on down-stream supervised financial text mining tasks which are Financial Sentence Boundary Detection (SBD), Financial Sentiment Analysis (SA) and Financial Question Answering (QA). For SBD task, FinBERT-large outperformed other models with mean F1-score 0.97 (mean F1-score for predicting beginning and ending tokens F1-scores separately), for SA task, again FinBERT-large outperformed other models with 0.93 F1 score and lastly for the QA task, FinBERT-large model outperformed other state-of-the-arts with 0.76 normalized discounted cumulative gain and 0.68 mean reciprocal rank which are two ranking evaluation measures used to evaluate financial question answering task.

Finance sector is a wide and inclusive field that includes many sub domains. Compli-

ance is one of the areas that can be categorized as a sub field of finance. In financial sector, actors such as banks, insurance or investment companies have to comply with some regulations for anti-money laundering and combating the financing of terrorism. To be able to act along these regulations, financial institutions need to monitor and evaluate customer operations, money flow and many parameters. So, mining the related data is very fruitful to that end. In financial compliance domain, in 2020, Jabbari et al. published a paper [56] where they proposed a french financial news corpus and annotation schema for NER along with the relation extraction task. Their expressions in their paper summarized the need and extensive usage areas of information extraction techniques in financial compliance field all over the world's financial institutions: *"Operating in fear of massive fines in case of undetected violation of the regulations, these (financial) institutes have set up a procedure called Know Your Customer (KYC). Today KYC procedure is a time consuming effort of collecting information through declarative questionnaires and watchlist lookups. Furthermore, each customer's profile need periodical reviews to ensure the accuracy of the provided information. The regulations in force also evolve over time and legislators do not specify a standard of sufficient information in order to prevent financial institutes from enforcing a bare minimum control. This led financial institutions to seek as much data as possible on their clients' financial activities. In this context, automated solutions can help financial institutions to continuously update the information that can impact the assessment of risk profiles, e.g. ownership and managerial changes and company's domains of activity."*.Authors presented an ontology of financial concepts and relations in the compliance sub domain. Additionally, they composed a French corpus which includes 130 financial news articles which were manually annotated in line with their proposed ontology. In their annotation schema, there are hierarchical connections between different entities. However, they did not annotate nested entities. For example, they annotated "Bank of England" as an organization but did not annotate "England" as a location. This annotation logic is in line with the logic adopted in FINTURK annotation schema where flat NER was taken into account. During the NER experiment, they used spaCy framework which allows for custom NER training and they trained spaCy's French model with their new entity types. They obtained 0.75355 and 0.87 as F1-scores of overall full and partial extractions, respectively.

Zhao et al.published a paper [57] where they proposed a RoBERTa based sentiment analysis and its key entity detection approach for online financial texts. They analyzed sentiment of text with RoBERTa and via NER or rule matching they got financial entity list and selected key entities from the text as entity list. Since NER detects all entities, it is not the right method for detecting key entities considering their importance and relevance to the target domain. Considering the task as sentence matching, after obtaining entities, they used RoBERTa model and entered each entity and financial text into the model to determine the key entity or entities. Model's predicted entities were said to be critical entities. After that, they selected the key entities with the tag, in other words the key entities which are most relevant to the given tags from the financial text. Authors gave "corruption" and "fraud" as example tags. This task was considered as a machine reading comprehension (MRC) task and they rewrited the financial texts as MRC articles. After, they provided new dataset to RoBERTa's MRC model and predicted answers were key entities. The distinctive side of this study is that; they were first using NER to detect entities and then modifying the data to another form and using MRC model detecting key entities for financial tags that they set forth in the beginning. With their RoBERTa-base model based technique they obtained 0.95258 and 0.85056 F1-scores as their best results with two datasets respectively which were better than BERT.

In [58], which is another NER study in financial domain, anonymized real life bank documents were used. Authors proposed a CRF-based NER system for NER of 10 different NE categories where 4 are generic (e-mail, time duration, organization, date) and remaining 6 are domain-specific (money amount, article number, abbreviation, law name, reference, ordinance) NEs. In overall, proposed system achieved 0.962 micro-averaged F1-score.

CHAPTER 3

METHOD

3.1 Background

Transformers gained popularity in Natural Language Processing domain and outpaced Convolutional Neural Networks(CNNs), Long Short Term Memory (LSTM) Networks and Recurrent Neural Networks(RNNs) in general. Transformer architecture was brought to the literature by Vaswani et al. [1] and became one of the most popular neural network based architectures. Transformers are composed of two divisions, encoder part receives the input and outputs a vectoral representation for the input where the decoder part uses the output of the encoder part and also the output of the entire system that was produced so far and using these two inputs, decoder part outputs the probabilities of each possible outcomes.

Transformers do not contain any recurrence and convolution mechanisms but they are still the most successful architectures in sequence to sequence problems. For instance, in an NLP problem, when we are working with especially longer sequences, if we are using RNNs as neural architectures, since these models are not so able to remember older information from the former tokens in a sequence, the former information can not be efficiently used when solving the problem. However, with a mechanism called "attention mechanism" used in Transformers, instead of attending just the most recent token, all of the sentence is taken into consideration when constituting the output. Input's representation is reinterpreted by the architecture using its encoder's subparts, namely self-attention mechanism and feed-forward network. With self-attention, the a word is encoded as a better representation by looking to the other words in that sequence. That is to say, during the process of deciding on the best representation

for a word in a sentence, other words and their positions are important for the model. Moreover, the outputs of self-attention layer are inputs for feed-forward layer. Thus, unlike the LSTMs or RNNs, Transformers have an important advantage in taking long-term dependencies into account which can be an important capability with long sequences in an NLP problem.

In addition, an important feature is also seen in Figure 3.1 which was taken from the original paper, during the calculation of output probabilities besides the input embedding, also the output embedding was also used which formed until then.

Since it is not necessary to feed just one input and receive one output at a time, Transformers are processing in a parallel manner and this increases efficiency compared to earlier architectures.

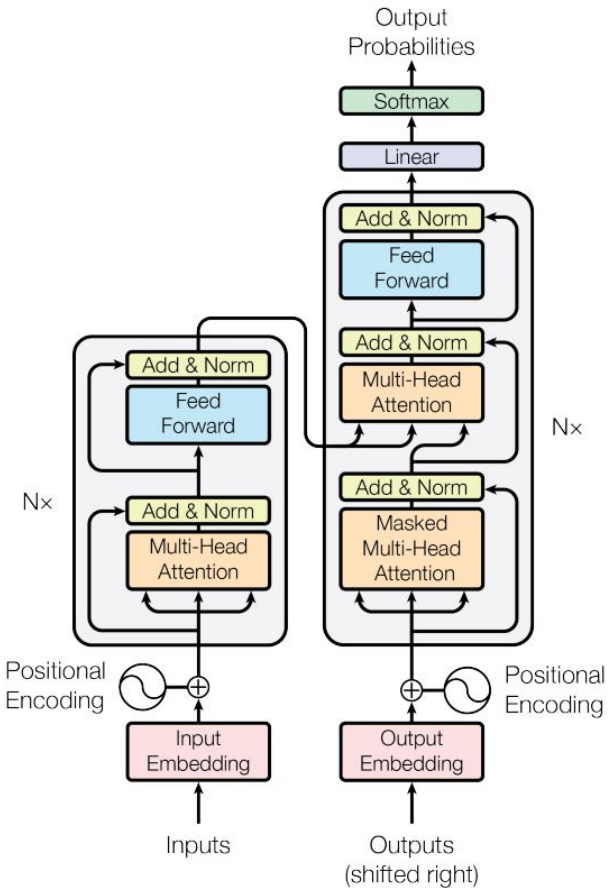


Figure 3.1: Transformer model architecture [1]

In addition, besides the input and output embeddings, also positional encodings of the

tokens are used in the architecture to gain information from the order of the tokens in a sequence.

In [59] Ezen-Can compared the performance of pre-trained BERT model (transformer-based architecture that we will mention in the next subsection) with the performance of a simple bidirectional LSTM model on intent classification using a small dataset. In the end, they concluded that *"bidirectional LSTM models can achieve significantly higher results than a BERT model for a small dataset and these simple models get trained in much less time than tuning the pre-trained counterparts."*It is worth stating that they had a really small corpus and added that the model performance is dependent on the task and the data. On the other hand, in [60] Hanslo compared performances of XLM-RoBERTa models with CRF and bi-LSTM models on NER task for low-resourced ten South African languages and concluded that XLM-RoBERTa models perfom better than other architectures.

In the scope of our thesis, we did not experimented with RNN or their variants LSTM networks we just used Transformer-based architectures since they proved their successes in NLP domain. However, if we had very small data to fine-tune, than it may be reasonable to experiment with RNN based smaller architectures as well.

In the subsections below, we provide general information and some key points about the architectures that we benefited from throughout our experiments. Deeper information about these architectures can be found in the original papers.

3.1.1 BERT

BERT which is a game changer architecture, "Bidirectional Encoder Representations from Transformers" was proposed by Devlin et al. [2] in 2018. BERT was pre-trained with two unsupervised tasks, masked language modeling (MLM) and Next Sentence Prediction (NSP).

BERT trains a transformer in a bidirectional way. In one directional methods, input sequence is read from left to right or vice versa. However, with this architecture, the words in the input sentence are read from left-to-right and from right-to-left at the same time. However,in traditional language modeling, the next word is being

predicted given the former words. To train a language model in a bidirectional way, we can train two models in opposite directions but in this case the model actually sees the word that it should predict. To tackle this problem, authors of BERT paper used MLM in pre-training and some of the tokens are masked and cannot be seen by the model. With the help of this technique, the model takes the surrounding words of a word into account when building a proper representation. Masked words are tried to be predicted correctly by the model looking at the surrounding words in the input. Thus, compared to the earlier techniques where a word's representation is the same vector independent from the surrounding words, BERT represents same word which appears in multi places, differently, depending on the context.

In another pre-training task of BERT, Next Sentence Prediction (NSP), two sentences are evaluated as being relevant or not. In other words, with this task used in the pre-training phase, model learns long term dependencies between sentences. In the training phase, MLM and NSP are also learnt by the model.

As input sequence, as authors stated in their paper as input token sequence, a single sentence or two sentences which are packed together can be thought. Packed sentences case is useful since BERT can be finetuned for downstream tasks where inputs like sentence pairs are observed such as Question Answering. In the scope of our thesis we studied on NER problem, which is another downstream task that BERT model can be fine-tuned (as seen in Figure 3.2 which was taken from the original paper) and in NER inputs are not sentence pairs but single sentences. Furthermore, BERT was designed for input sequences up to 512 tokens and we took this maximum sequence number into account during our experiments.

To represent input, input embeddings are constructed using summation of token embeddings, segmentation embeddings and position embeddings. During the token embeddings construction, each token is represented as a vector of fixed size using the WordPiece method.Segment embeddings are about the word's (token's) position at sentence level (considering the case that input consists of pair of sentences). Besides, it is important to note that in NER, there are no sentence pairs as inputs. In addition, position embeddings are about the token's position in that sentence. In BERT, absolute position embedding is used namely, absolute positions of words in a sentence are

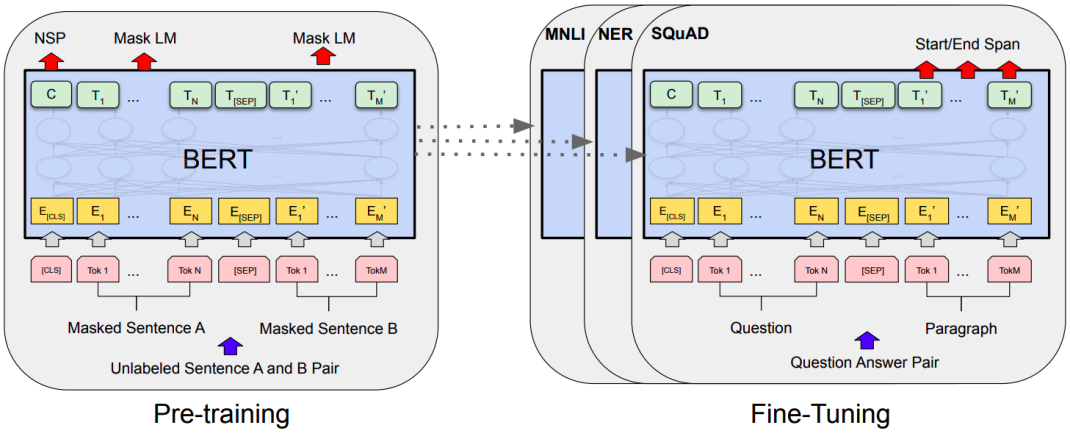


Figure 3.2: BERT Pre-training and Fine-tuning [2]

taken into account rather than the relative positions of words considering each other.

BERT has 2 versions BERT-Base and BERT-Large with 12 and 24 layers, 110 million and 345 million parameters, respectively. BERT model was pre-trained using large amount of unlabeled data from different resources over MLM and NSP pre-training tasks. After pre-training, using the relevant labeled data, BERT was fine-tuned for different downstream tasks including NER.

Lastly; it is an advantage that BERT is an open-source model which is easy to access from platforms like GitHub or Hugging Face. Besides the English and multilingual versions of BERT, there are many monolingual BERT models which were trained for different languages. In our study, we also used the BERT model trained for Turkish besides the original model.

3.1.2 DistilBERT

After BERT, many BERT-like architectures were proposed. One of them is DistilBERT by [3]. DistilBERT was an improvement over BERT in terms of the inference speed. Authors mentioned that increasing computational and memory requirements, more successful but bigger models are proposed as time passes. In Figure 3.3 authors visually presented the number of parameters of different popular models which proposed in a few years and it is obvious that DistilBERT is an outlier with regards to the

trend of proposing heavy models with high number of parameters.

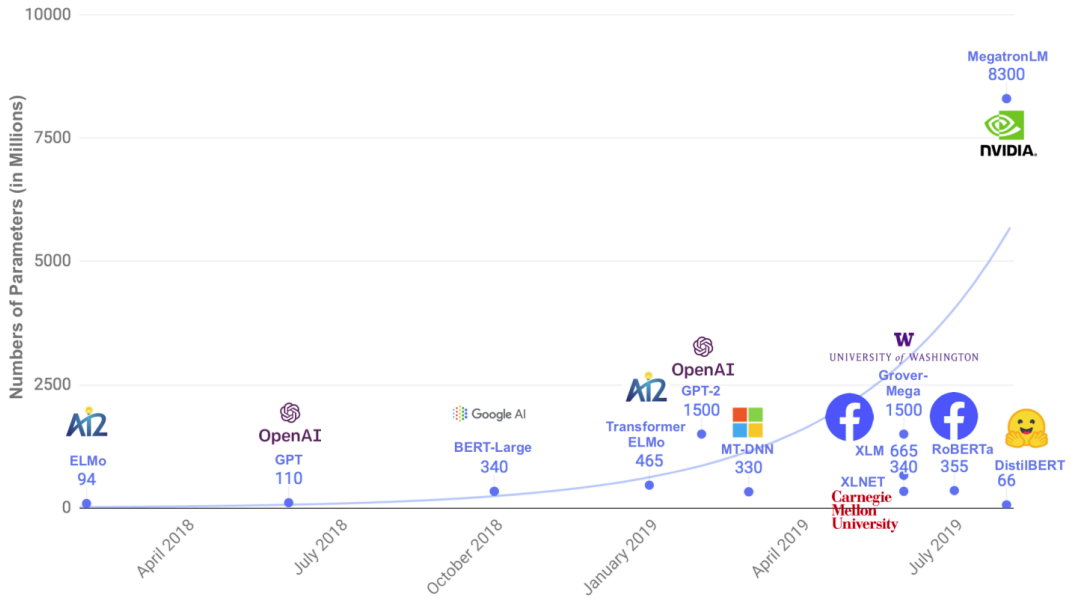


Figure 3.3: Parameter counts of several pre-trained language models [3]

Using the distillation process, they compressed BERT to a smaller model. They also emphasized that it is possible to reach closer performances to larger models in less inference time and computational training budget when using lighter models. During distillation, DistilBERT as a student model, was trained under the supervision of BERT which was the teacher model. According to the paper, they removed token-type-embeddings and pooler, comparing DistilBERT to BERT and they also did not use the NSP objective in training. At the end, researchers obtained a faster, lighter and cheaper model with less number of layers than BERT. Moreover, as the authors present as result of their experiments, DistilBERT has 66 million parameters and 410 seconds as the inference time whereas BERT Base model has 110 million parameters and 668 seconds as the inference time which is %60 slower than DistilBERT.

Similar to BERT, it is an advantage that DistilBERT and its versions pre-trained for various languages are publicly available on the web. We experimented with both DistilBERT Base Multilingual Cased model and DistilBERT Base Turkish Cased model with 4 datasets we have.

3.1.3 XLM-RoBERTa

In 2019, XLM-RoBERTa model was published by Conneau et al. [61] as an improvement over the former method XLM [4] which stands for cross-lingual language modeling and is also proposed as a transformer based multilingual language model. Actually, after the success of BERT, the main objective of XLM models was to propose improvements on cross-lingual modeling. Since more than one languages are in question, authors processed all languages with Byte Pair Encoding technique to obtain a shared subword vocabulary for these different languages. In the XLM training setting, authors proposed three different pre-training objectives. In Causal Language Modeling (CLM), the probability of a word is determined via looking at the previous words in that sentence but this task cannot be scaled to cross-lingual setting. MLM is the same training objective as in BERT. CLM and MLM objectives require just monolingual data. The Translation Language Modeling (TLM) objective is used where each training set element is composed of parallel texts composed by concatenating texts in different languages and since the MLM was utilized by masking some random tokens from each text in different languages, it is possible for the model to use the surrounding information of masked tokens in both texts in both languages when predicting masked tokens. Thus, compared to BERT, where only a monolingual text is available as input, now the model has much more opportunity for using information to benefit from in prediction. Researchers trained the model using just MLM or TLM tasks and also with both tasks as well.

Furthermore, in XLM, authors used data from 15 different languages but afterwards, using much more training data amount (more than 2 terabytes of CommomCrawl data) compared to XLM, Facebook AI Team used a model containing 24 layers, 1024 hidden states and trained XLM-RoBERTa model on 100 languages including Turkish. In addition, there is a different technique for subsword tokenization and there is no language embedding, model has to determine which language it is dealing with by its own. Actually, XLM-RoBERTa is not very different from multilingual BERT model, the pre-training data was different from BERT, the architecture is similar but the authors made important changes in the pre-training process which were used formerly in RoBERTa. Therefore, the "RoBERTa" part of the model comes from having

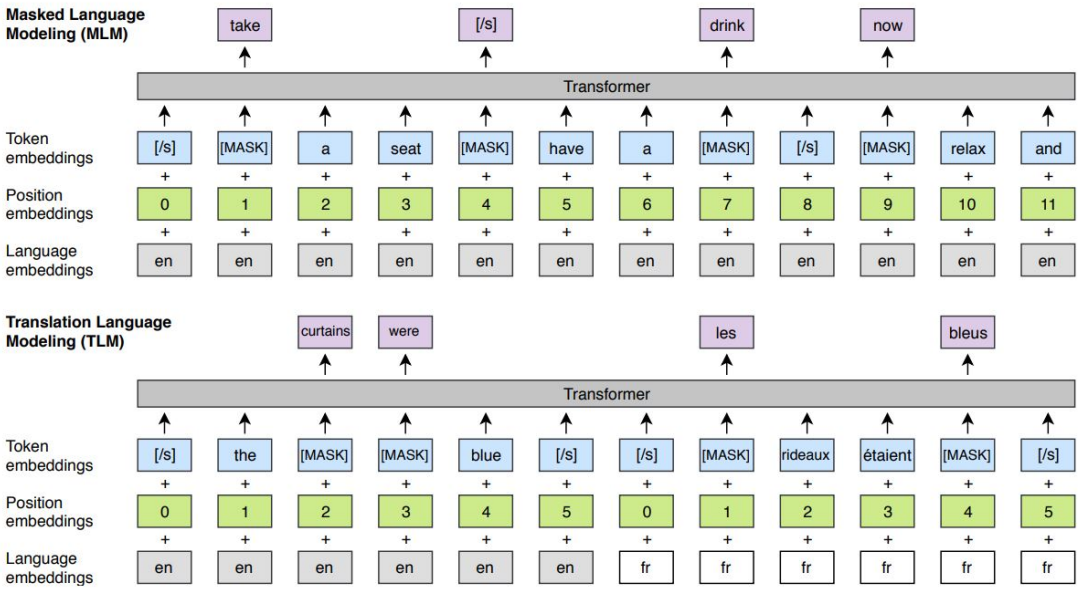


Figure 3.4: Cross-lingual language model pre-training [4]

the same training task (just MLM and not NSP) with RoBERTa[62]. RoBERTa was proposed in 2019 since the authors thought that BERT was undertrained and there are rooms for improvement. In RoBERTa, researchers replaced the static masking in BERT with dynamic masking, removed the NSP task, used more training data than BERT, trained on longer sequences. If we move back to XLM-RoBERTa, for NER task, XLM-RoBERTa outperformed multilingual BERT with +2.4%F1-score. It is also stated that the model performs well for also low resourced languages. Resources also report that their best model XLM-RoBERTa outperformed multilingual BERT model on cross-lingual classification by up to 23% accuracy on low-resource languages.

XLM-RoBERTa model is publicly available on many platforms, throughout our experiments we benefited from the model on Hugging Face platform.

3.1.4 ELECTRA

In 2020, Clark et al. [5] presented an alternative transformer based architecture, namely ELECTRA which uses Replaced Token Detection (RTD) mechanism as the pre-training task which can be understood as the small MLM mechanism that was

used in BERT architecture. In ELECTRA architecture, there exist a generator and a discriminator network as seen in Figure 3.5 which was taken from the original paper.

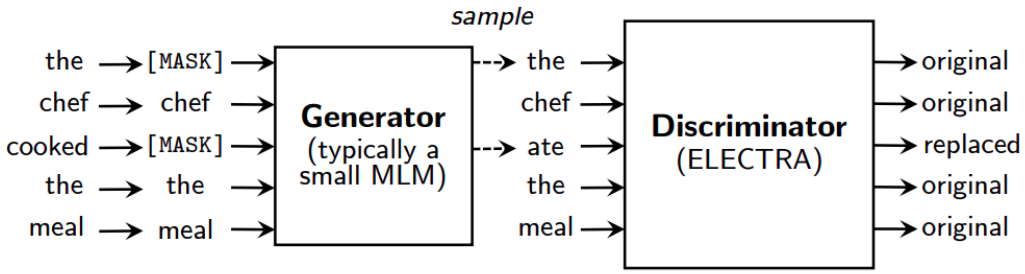


Figure 3.5: ELECTRA model Replaced Token Detection Mechanism Illustration [5]

Authors stated that, in RTD, they corrupted the input via changing them using the sample outputs (as potential inputs) of a small generator network. Moreover, they trained a discriminative model which predicts whether each token in the corrupted input was replaced by a generator sample or not. In the paper, it is also mentioned that this RTD mechanism is more efficient than MLM mechanism since it is dealing with all input tokens instead of only masked tokens in MLM mechanism. In other words, in BERT with MLM, only masked tokens are trying to be predicted but with RTD mechanism, all of the masked or unmasked tokens are examined by the discriminator network to find out whether they are replaced or not. In addition, after the pre-training phase the generator network was not used and only the discriminator network was fine-tuned for the on downstream tasks.

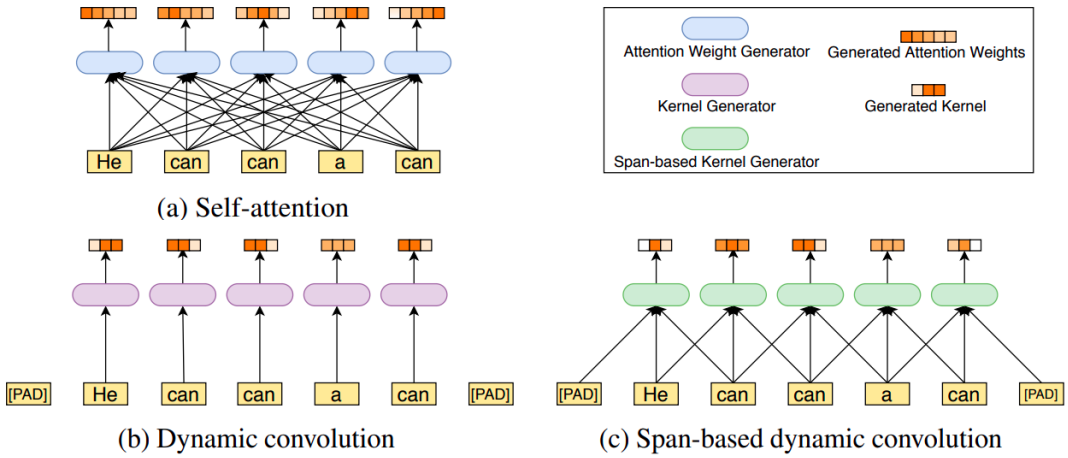
Various pre-trained versions of ELECTRA Base Cased Discriminator model are available publicly. In our experiments, we also experimented with the ELECTRA Base Turkish Cased Discriminator model where we train, evaluate and predict using our datasets, in other words we fine-tuned the discriminator model using our dataset and new labels.

In the paper it is also mentioned that ELECTRA outperforms BERT, RoBERTa and XLNet in learning contextual representations.

3.1.5 ConvBERT

Another neural architecture which was presented in 2020 is ConvBERT[6]. Authors mentioned that BERT has large memory requirements and computation cost since it benefits from global self-attention. Although all of the heads are global heads, some of the heads need to learn local dependencies and this brings a computational redundancy in BERT. In this architecture, researchers replaced some of the attention heads that were used to learn global dependencies in BERT,by different convolution-based mechanisms to learn local dependencies.

Chen et al.[63] presented the dynamic convolution and stated:*"Dynamic Convolution, a new design that increases model complexity without increasing the network depth or width. Instead of using a single convolution kernel per layer, dynamic convolution aggregates multiple parallel convolution kernels dynamically based upon their attentions, which are input dependent. Assembling multiple kernels is not only computationally efficient due to the small kernel size, but also has more representation power since these kernels are aggregated in a non-linear way via attention."*. Moreover, Jiang et al. [6] presented new span-based convolution mechanism. Self-attention, dynamic convolution and span-based dynamic convolution mechanisms are illustrated in the Figure 3.6 which was taken from the original paper.



(a) Self-attention

(b) Dynamic convolution

(c) Span-based dynamic convolution

Figure 3.6: Self attention, dynamic convolution and span-based dynamic convolution [6]

To be more specific, span-based dynamic convolution mechanism was used in ConvBERT since it was aimed to perform local self-attention. As explained in detail in the paper, in the span-based dynamic convolution, input is a span of tokens and mechanism produces more adaptive kernel generators to enable the discrimination of the generated kernels for same tokens within different context. As being an improvement over BERT, ConvBERT is also an improvement over ELECTRA since it was pretrained on the RTD task that was used in ELECTRA. In addition, authors also stated that ConvBERT Base model outperformed BERT and ELECTRA-Base models in various downstream tasks.

Various pre-trained versions of ConvBERT model are available publicly. In the scope of our study, we used ConvBERT Base Turkish Cased model.

3.1.6 DeBERTa

DeBERTa[64] (Decoding enhanced BERT with disentangled attention) is another architecture proposed by authors from Microsoft in 2020. DeBERTa improves BERT and RoBERTa introducing "disentangled attention mechanism" and "enhanced mask decoder" techniques. In disentangled attention mechanism technique, word representation is based on its content and position separately as opposed to BERT architecture where the content and position information are combined to obtain representation. DeBERTa is pre-trained used MLM used in BERT. Moreover, in disentangled attention mechanism, relative positions of words are taken into account since the strengths of dependencies between words depend on their positions in a sentence or different sentences, relative to each other. However, researchers also mentioned that, absolute positions of words also matter for instance, considering their syntactic roles in sentences. Thus, in enhanced mask decoder, words are represented with an aggregation of their content, relative and absolute position information. DeBERTa outperformed BERT, RoBERTa, XLNet and ELECTRA architectures in many tasks given in detail throughout the paper. The Figure3.7 which illustrates the architecture of the model was taken from the related website of the company.

After proposing DeBERTa, DeBERTaV3[65] was presented as an improved version of DeBERTa which benefits from ELECTRA-style pretraining with gradient- disen-

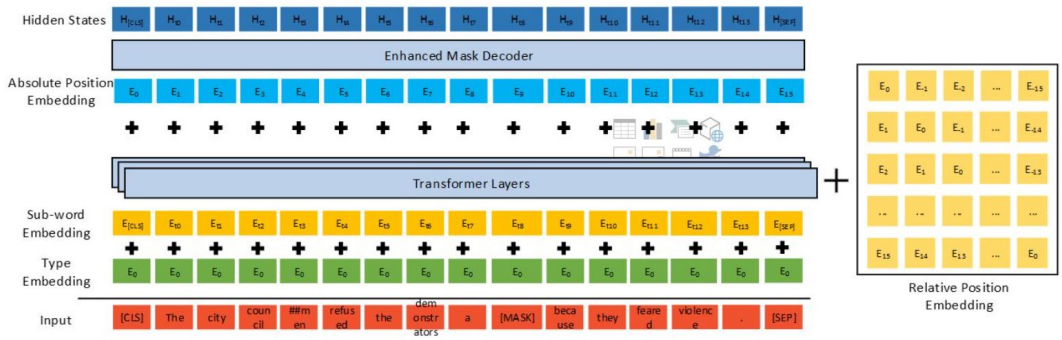


Figure 3.7: DeBERTa Architecture [7]

tangled embedding sharing. Researchers replaced the Masked Language Modeling task in DeBERTa with Replaced Token Detection task where a discriminator network is used to predict if the input token is original token or if it is replaced version by the generator network. In ELECTRA, where same token embeddings are used in generator and discriminator networks, the token embeddings are influenced by training losses from different networks (i.e. generator and discriminator networks) in two different directions and this brings an inefficiency in training phase. To solve this problem, researchers used a gradient-disentangled embedding sharing method where the embeddings of generator are shared with discriminator but discriminator gradients are not backpropagated to update embeddings of generator network. mDeBERTaV3 variants outperformed BERT, RoBERTa, XLNet and ELECTRA in various NLP tasks as it stated in the paper.

mDeBERTa model is publicly available and we used it from Hugging Face platform.

In the scope of this thesis, as it can be seen in Chapter 4, we applied these models on different datasets where two of them are newly proposed by us and remaining two are already existed in literature.

3.2 Dataset

In the scope of this thesis, mainly 4 different datasets are used where 2 of these datasets were used in other studies formerly and remaining 2 are newly constructed for this study.

Details of existed and formerly used datasets are given in 4.

3.2.1 Preparation of FINTURK Dataset

FINTURK datasets were constructed via preprocessing and annotating the raw news text from Dünya Newspaper. Annotated portion of the text is from January, February and March news of 2015.

FINTURK datasets contain approximately 500.000 (502.657) tokens and has 4 versions as seen in the table 3.1. Since we also used other datasets that were used in different studies before, to achieve the compatibility considering annotation format, FINTURK datasets, all versions (parent and children versions) were annotated with BIO and NON-BIO (RAW) format. It is important to note that; in the literature, there are many BIO-type representations which exhibit small or big differences and some of them seem to be different formats when just looking to their names but they are basically the same annotation format. For example: IOB2 format is commonly referred to as IOB, BIO or CoNLL format [9].

Detailed explanations about each named entity type can be seen in further subsections of this section.

3.2.2 Preprocessing

Dünya is a newspaper which generally presents news about finance, general world and country agenda, important events and multiple business domains such as tourism, automotive and agriculture.

In the preprocessing step, first of all the news texts was reviewed and generally, news that are related to financial domain and general business domains were kept. After this reviewing step, tokenization step followed. For tokenization, first of all lines of news texts were splitted by whitespaces. Afterwards, Python Natural Language Toolkit (NLTK) package [66] was used and with the nltk.tokenize.WordPunctTokenizer() method, tokens were extracted from the splitted lines. There are two examples of tokenized sentences from datasets in Table3.2. As seen from the examples, words and

Table 3.1: FINTURK500K Dataset Versions

FINTURK500K Dataset Version	**Information About the Dataset Version**
FINTURK500K PARENTS BIO	Contains tokens tagged according to the BIO tagging schema. Here, only the parent tags were used for annotation. For example, financial and non-financial organization entities were both tagged as organization entity. Furhermore, an organization entity was tagged like: "B-ORG I-ORG I-ORG".
FINTURK500K PARENTS NON-BIO	This is the NON-BIO tagged version of the version explained above. For example an organization entity was tagged like: ORG ORG ORG". The reason for using this NON-BIO schema for annotation is to be able to measure the effect of annotation schema on performance and to compare the performance of the experiments conducted with FINTURK500K dataset versions with TUR3 and EXTEND7 datasets which were annotated using NON-BIO tagging schema.
FINTURK500K CHILDREN BIO	Contains tokens tagged accoring to the BIO tagging schema. Here, children tags were used for annotation. For example, a financial organization was tagged like: "B-FIN-ORG I-FIN-ORG I-FIN-ORG" and a non-financial organization was tagged like: "B-NON-FIN-ORG I-NON-FIN-ORG I-NON-FIN-ORG".
FINTURK500K CHILDREN NON-BIO	This version of the dataset is constructed with NON-BIO tagging schema. Similar to the FINTURK500K-PARENTS-NON-BIO version this dataset was prepared for the same comparison reasons.

affixes separated from word by apostrophes were counted as separate tokens. This may help the model to learn proper nouns such as person, organization or location names. Also, punctuations such as "." were counted as separate tokens as well.

Moreover, in both of the FINTURK dataset versions, an input token sequence is a sentence and sentences are separated by blank lines. Also, in each line a word (token) has the label (NE tag) next to it separated by a tab.

Table 3.2: Examples of the Tokenization of Sentences

Token	**Parent Tag (NON-BIO)**	**Child Tag (NON-BIO)**
4.69	MONETARY-VALUE	MONETARY-VALUE
lira	CURRENCY	CURRENCY
yerine	O	O
4.74	MONETARY-VALUE	MONETARY-VALUE
lira	CURRENCY	CURRENCY
vergi	O	O
ödeyecekler	O	O
.	O	O

Token	**Parent Tag (BIO)**	**Child Tag (BIO)**
Ekonomi	B-ORG	B-FIN-ORG
Bakanlığı	I-ORG	I-FIN-ORG
susamın	B-COMMODITY	B-AGRIC-COMMODITY
gümrük	B-TAX	B-TAX
vergisini	I-TAX	I-TAX
yüzde	B-PERCENT	B-PERCENT
56	I-PERCENT	I-PERCENT
düşürdü	O	O
.	O	O

Token	**Parent Tag (NON-BIO)**	**Child Tag (NON-BIO)**
Draghi	PERSON	PERSON
:	O	O
ECB	ORG	FIN-ORG
’	O	O
nin	O	O
fiyat	O	O
istikrarını	O	O
sağlaması	O	O
zorlaştı	O	O
.	O	O

3.2.3 Annotation Procedure

In 1998, 7th of the Message Understanding Conferences (MUC) was held which is an event sponsored by DARPA (Defense Advanced Research Projects Agency). These series of conferences (MUC-1 to MUC-7) aimed to improve information extraction studies basically. In MUC-7, the details of named entity recognition task are specified and a detailed guideline for annotating named entities which expose different characteristics was provided [13]. In this study, named entities (ENAMEX), temporal expressions (TIMEX) and number expressions (NUMEX) were defined as in Table 3.3. In addition, in the scope of this thesis, ENAMEX, NUMEX and TIMEX entities are named as NEs in general.

Table 3.3: MUC-7 Named Entity Types and Corresponding Definitions

Type of Named Entity	**Upper Tag Category**	**Definition**
ORGANIZATION	**ENAMEX**	Named corporate, governmental, or other organizational entity (limited to proper names, acronyms, miscellaneous other unique identifiers)
PERSON	**ENAMEX**	Named person or family (limited to proper names, acronyms, miscellaneous other unique identifiers)
LOCATION	**ENAMEX**	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains etc.) (limited to proper names, acronyms, miscellaneous other unique identifiers)
DATE	**TIMEX**	Complete or partial date expression (both absolute and relative)
TIME	**TIMEX**	Complete or partial expression of time of day (both absolute and relative)
MONEY	**NUMEX**	Monetary expression (either numeric or alphabetic form)
PERCENTAGE	**NUMEX**	Percentage (either numeric or alphabetic form))

In this very detailed work[13] there is also a specific chapter dedicated to guidelines for markup of exceptional constructions and in appendices more detailed rules are provided for ENAMEX, TIMEX and NUMEX, with examples.

For FINTURK datasets, a similar detailed guideline is tried to be provided in the following sub chapter along with some comparisons with similarities and differences between MUC-7's guideline.

3.2.4 Detailed Explanations for the Annotation of Each Named Entity Type

3.2.4.1 Person

Name of a person. In the MUC-7 guideline[13], it is said that family names and other proper names (like full names) are considered as PERSON named entity. In FINTURK annotation, from the context, if the first name (without second name) of a person is enough for the annotator to understand who it designates from the context, it is annotated as PERSON. Also, like in MUC-7 guideline, acronyms and other proper names are under this category. It is important to note that; titles like job titles or academic degrees are not tagged as a part of the PERSON named entity.

3.2.4.2 Organization

Name of any kind of organizational entity. This entity may be a corporate or a civil society organization. Like in MUC-7 guideline[13], if the entity does not contain the full name of the organization like "....Inc." (or "...A.Ş." in Turkish) if it is clear from the context to the annotator then it is annotated as organization. For example, "Reuters" is annotated as an organization since it is clear that it refers to "Thomson Reuters Corporation".

In MUC-7 guideline[13], it is stated that nick name aliases for organization names are tagged as organization entities. However, in FINTURK annotation, nick names are not annotated as organization.

Moreover, in MUC7 guideline[13], in Appendix A.1.5. there is a subchapter dedicated to entity -expressions that "possess" other entity-expressions where all distinct expression is tagged seperately. However, in FINTURK annotation, if an organization name contains another organization name in it, such as "Oflaz Şirketler Grubu Yönetim Kurulu" (Oflaz Company Union Board of Directors) all of the entity is tagged as a single organization, namely, "Yönetim Kurulu" (Board of Directors) part is not tagged separately as a new organization. As another example, other than annotating "Uluslararası İlişkiler Bölümü" (Department of International Relations) separately as another organization, "Marmara Üniversitesi Uluslararası İlişkiler Bölümü" (Marmara

University Department of International Relations) is annotated as a single organization entity even if the "Uluslararası İlişkiler Bölümü"(Department of International Relations) is another organization which has an autonomous government itself.

Furthermore, as the motivation of this dataset construction study implies, besides the classical well-known "ORGANIZATION" named entity, there are both financial and non-financial organization annotations. There is no organization other than these two groups.

Another hue in the annotation of organizations is that the case where the abbreviation of the name of the organization is given in paranthesis after the long name of the organization. "Uluslararası Para Fonu (IMF)" (International Monetary Fund) is tagged fully as a single organization entity including the "(IMF)" abbreviation. However, if there is any single affix between the long name and the abbreviation, then the abbreviation is tagged as another organization itself. To put it in different way, a part of the sentence like "Uluslararası Para Fonu'na (IMF)" (to the International Monetary Fund) is annotated separately, "Uluslararası Para Fonu" is tagged as an organization, and "IMF" is tagged another organization.

Similar to the MUC-7 guideline[13], if a token or a group of tokens represent both an organization and a location, such as an airport, an exposition center, if from the context annotator understands that a location is in question, then, tokens are annotated as location.

According to MUC-7 guideline[13], it is said that miscellaneous types of proper names are not tagged as enamex. As an example, expression "Wall Street Journal" is given as a non-tagged expression. However, this expression is tagged as an organization in FINTURK datasets if it is clear from the context that stands for an organization.

In MUC-7 guideline[13] it is also stated that plural names that do not identify a single entity however, such group of entities are completely tagged as an entity in FINTURK if every single one of them are under the same NE category.

Financial Organization

Financial organizations are organizations that take actions in banking, insurance, economic research fields but not limited to these forms of corporations. This tag also includes associations in finance and economy fields and other very-related domains. Examples below show these type of organizations more clearly.

“Emeklilik Gözetim Merkezi (EGM)” (Pension Monitoring Center) is tagged as a financial organization fully.

“Euro Bölgesi” (Euro Zone) is tagged fully as a financial organization since it is clear to the annotator that it refers to the “Euro Alanı” (Euro Area) which is the official name of the monetary union.

“Sosyal Güvenlik Kurumu Başkanlığı (SGK)” (Directorate of Social Security Institution) is fully annotated as financial organization since social security is considered in scope of insurance which is a financial-related domain.

Non-Financial Organization

Non-financial organizations are the ones which are not financial organizations to say the least. If an organization is not very close to financial domain, for example a consultancy firm conducts financial consultation besides consultations in other fields, and it is annotated as a non-financial organization in FINTURK annotation schema as in the following example.

“Deloitte Türkiye” (Deloitte Turkey) is tagged as a non-financial organization even if from the context it is clear that the consultation is about the financial domain for that news text. In other words, the main area of activity is taken into consideration.

3.2.4.3 Location

In MUC-7 guideline[13], “location” is defined as “name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)”, which also applies to FINTURK datasets’ annotation schema.

As mentioned in the "ORGANIZATION" section, if a token or a group of tokens represent both an organization name and a location name, annotation is done by the understanding from the context. For example, in "...Adana'nın Ceyhan ilçesindeki Haydar Aliyev Deniz Terminali'nde tankerlere petrol yüklemesi sürüyor." (Oil loading continues to tankers at Haydar Aliyev Marine Terminal in Adana's Ceyhan district.), the expression "Haydar Aliyev Deniz Terminali" (Haydar Aliyev Marine Terminal) is annotated as a location even if it also means an sub organization of BOTAŞ International organization at the same time. Abbreviations of location names such as "ABD" (USA) are also tagged as location if it is clear from the context. Unlike the MUC-7 guideline[13], nick name alias for locations are not tagged as location entities in FINTURK annotation.

Furthermore, similar to the organization type named entities, if an expression possesses another expression where both indicate locations by just theirselves, all of this expression is tagged as a single location entity in FINTURK, unlike to the MUC-7 guideline[13]. For example, in "...İzmir'in Bayraklı ilçesinde her cumartesi kurulan semt pazarındaki..." (...in the neighborhood market, which is held every Saturday in the Bayraklı district of İzmir...), the expression "İzmir'in Bayraklı ilçesinde" (in Bayraklı district of İzmir) is tagged as a single location entity in FINTURK datasets even if there are actually two different location entities ("İzmir" and "Bayraklı ilçesi") according to the MUC-7 guideline.

3.2.4.4 Monetary Value

Similar to MUC-7 guideline[13], these are monetary expressions either in a numeric or an alphabetic form. However in FINTURK annotation, expressions that refer to a currency are tagged as another named entity type, "CURRENCY".

Similar to MUC-7 guideline[13], approximations are not considered in the scope of the monetary value tag. For example, in the sentence: "Bunun ekonomize katkısı 300 milyon lira civarıdır." (Its contribution to our economy is around 300 million TL), "civarıdır" (around) token is not annotated as a continuation of monetary value tag for "300 milyon lira" (300 million TL), it is annotated as "other".

Moreover, parities of different currency types are annotated as monetary values since they express a monetary amount even if parity is a ratio. For example, for "...kazanmasının ardından euro/dolar paritesi 1.12 düzeyinin de altına inerek..." (...after winning euro /dollar parity falling below the 1.12 level...) expression, "1.12" is tagged as a monetary value.

For multipliers, in MUC-7 guideline[13] it is said that "modifiers that indicate the multiplied value of a number unit should be included in the tagged string, if the modifier is a substitute for a specific digit (or the indefinite article or other quantitative determiner) within the monetary or percentage expression." For FINTURK, same is applied. For example, for the sentence "Türk ihracatçıların Rusya'dan yüzlerce milyon dolarlık gecikmiş alacağı bulunuyor." (Turkish exporters have hundreds of millions of dollars of overdue receivables from Russia.), expression "yüzlerce milyon dolarlık" (hundreds of millions of dollars) is tagged as a single monetary value entity.

3.2.4.5 Currency

"Currency" named entity tag represents units of various currencies over the world. In FINTURK's annotation, all of the currency types are annotated as a separate tag other than monetary value, different from MUC-7 annotation[13].

Besides, according to MUC-7 guideline[13], "Juxtaposed strings expressing monetary values in two different currencies are to be tagged separately. Same applies for FINTURK annotation. For an expression "... 27 milyar euro (30 milyar dolar)" (27 billion euros (30 billion dollars)), "27 milyar euro" and "30 billion dollars" are annotated as separate monetary values followed by currency tags for both "euro" and "dolar" expressions.

3.2.4.6 Percentage

This category expresses percentage which can be in both numeric or non-numeric form. Expressions that represent an interval between two percentages are completely annotated as a single percentage tag. For example: for a part of sentence like " Dolayısıyla aslında biz sigarada yüzde 5 ile 9 arasında bir fiyat artışına gittik ..."

(Therefore, we actually went for a price increase of between 5 and 9 percent in cigarettes...), expression "yüzde 5 ile 9 arasında" (between 5 and 9 percent) is annotated as a single percentage named entity.

Furthermore, in case of approximators, as stated in the MUC-7 guideline[13], "modifying words that indicate the approximate value of a number or a "relative position" to a number are generally to be excluded" in FINTURK. For instance in "Yüzde 70'e varan indirim" (Up to 70% discount) expression, only "Yüzde 70" (70%) is tagged as a percentage entity, the expression, namely "'e varan" which shows an approximation is ignored and tagged as "other". Similarly, for "minimum yüzde 25" (minimum 25%) expression "minimum" (minimum) is not tagged as percentage, "yüzde 25" (25%) is tagged as a single percentage named entity.

Similar to MUC7 guideline[13], "minus" expression ("eksi" in Turkish) is tagged in the scope of the relevant percentage annotation. For example, in the expression "...mevduat faizini yüzde eksi 0.75'e düşürmesinin..." (...reducing the deposit rate to minus 0.75 percent...) "eksi" (minus) is tagged as a single percentage entity with "0.75".

In MUC-7 guideline[13], as it is stated "values that do not use percentage terms to indicate percentages are not to be tagged". In FINTURK, same applies. For example, for a sub sentence like: "...Türkiye'ye gelen turistlerin üçte birinin Antalya'yı ziyaret ettiğine değinen Türker..." (...pointing out that one third of the tourists coming to Turkey visit Antalya, Türker...), "üçte biri" (one third) expresses a percentage mathematically, but is not tagged as a percentage entity in FINTURK.

3.2.4.7 Date

According to MUC-7 guideline[13], "An expression of financial quarters or halves of the year must indicate which quarter or half, such as "fourth quarter", "first half". Note that there are no proper names, per se, representing these time periods. Nonetheless, these types of time expressions are important in the business domain and are therefore to be tagged." In financial domain, quarters of years have special importance. For example, banks, corporations quoted on the stock exchange and other

similar organizations announce their profits quarterly and there are many more activities arranged quarterly in financial sector. Thus, in FINTURK, these expressions aforementioned are annotated as "DATE" entities.

In addition, similar to MUC-7 guideline[13], expressions indicating a decade, or a group of years are annotated as date entities. For example, for a sentence like “1990’larda ABD ekonomisi yıllık ortalama yüzde 3,4 büyüyordu.” (In the 1990s, the US economy was growing at an average of 3.4 percent annually.), “1990’larda” (In the 1990s) expression is tagged as a single date entity.

According to the MUC-7 guideline[13], “Words or phrases modifying expressions (such as “around” or “about”) will not be tagged. Only the actual temporal expression itself is to be tagged.”, the situation is same for FINTURK datasets’ annotation rules.

Unlike the MUC-7 guideline[13], relative temporal expressions are not tagged in FINTURK. In MUC-7 guideline it is stated that expressions like “yesterday”, “today”, “evening”, “afternoon”, “this month” are tagged as date named entities.

3.2.4.8 Time

TIME named entity stands for temporal tokens that express a time smaller than a day.Similar to MUC-7 guideline[13], if a time entity is just after a date entity, they are tagged separately with proper tags. For example, for a sentence part such as “...İstanbul Kongre Merkezi’nde 9 Şubat Pazartesi saat 15.00’te düzenlenecek.” (...it will be held at Istanbul Congress Center on Monday, February 9 at 15:00.), “9 Şubat Pazartesi” (Monday, February 9) is tagged as a single date entity whereas “saat 15.00’te” (at 15.00) is tagged as a single time entity.

As stated in the previous section, expressions for relative temporal situations are not tagged in FINTURK datasets, that applies to time entities.

3.2.4.9 Index

"Index" named entity tag stands for any index type such as stock market indices, price indices, real estate indices and so on. Indices are important indicators in financial world, economic parameters may be affected by changes in related indices and they provide information about the course of financial events. Thus, index named entity is another type that took place in FINTURK annotation.

Some examples to this label from the FINTURK datasets are, "Dow Jones endeksi" (Dow Jones Index), "Standard&Poor's 500 endeksi" (Standard&Poor's 500 Index), "Nasdaq Teknoloji endeksi" (Nasdaq Technology Index), "ASE endeksi" (ASE Index).

In FINTURK, the world "endeks" (index) is tagged within the same index named entity tag with the previous word. For example, for the expression "Ücretliler Geçinme Endeksi" (Wage earners cost of living index) all of the tokens are tagged as a single index entity.

3.2.4.10 Business-Sector

As its name implies this is the category for various business sectors such as "financial sector", "automotive sector", "industry" etc. Although it has such a simple definition, the annotation process of this named entity is a bit complicated. Some expressions may express both a sector name or an asset name related to that sector depending on the context. For example, word "konut" (housing) both is an asset and is the name of a business sector. As another similar instance; word "inşaat" (building construction) is both an act and name of the related sector. In such situations annotation is completed considering the meaning that the annotator understands from the contexts,however the main tendency is to annotate expressions which express a generality for something throughout that sector, as the business sector. Instances that indicate a kind of product in that sector (such as buildings in housing sector) are not tagged as a business sector.

The financial news corpus is very rich in terms of different news related to different

sectors and their sub sectors. Additionally, there may be multiple names used for a specific sector such in the case of "emlak", “konut” or “gayrimenkul” which are used for “housing” sector. In such cases, any instance of these names are tagged as a business sector entity.

Another thing to mention about this category is that; if a token expresses both a commodity and a business sector, it is tagged as a commodity. For instance, “süt üretimi” (milk production) is a business sector but also contains “süt” (milk), an agricultural commodity entity in it. Therefore, “süt” (milk) is tagged as an agricultural commodity where “üretimi” (production) is tagged as “other”.

3.2.4.11 Commodity

As a definition, a commodity is a basic good that is interchanged with other goods of the same type in commerce. In FINTURK, there are agricultural commodities, natural resources and other commodities under this category. Commodities that are neither agricultural commodities nor natural resources are tagged as just “commodity”.

Agricultural Commodity

Agricultural commodities are agroproducts that are commodities. Examples include crops and animals raised in farms. In FINTURK annotation, secondary agricultural commodities that are made from primary agricultural commodities via a somehow complex process, are not count as agricultural commodities. To be more specific, for example, “milk” is an agricultural commodity which is directly obtained from cows, which is another agricultural commodity, and that is not a very complex process. However, “cheese” which is obtained from “milk” through some processes via humans is not counted as an agricultural commodity in FINTURK annotation schema. The reason for this simplification is to make annotation easier and more coherent. Examples to these types of commodities from FINTURK are as follows:

- For the sentence “...Buğdayda 2015 sezonunun olumlu geçeceğine inanıyoruz.” (We believe that the 2015 season will be positive in wheat.), the word “wheat” is tagged as an agricultural commodity.

- For the sentence "Buna göre,ocakta toplanan inek sütü miktarı, geçen yılın aynı ayına göre yüzde 0,8 artarak 714 bin 510 ton oldu." (Accordingly, the amount of cow's milk collected in January increased by 0.8 percent compared to the same month of the previous year and became 714 thousand 510 tons.), the expression "inek sütü" (cow's milk) is tagged as a single agricultural commodity.

- For the sentence "Mutfağın zam şampiyonu kuru kayısı." (The price champion of the kitchen is dried apricots), the expression "kuru kayısı" (dried apricots) is tagged as an agricultural commodity.

Natural Resource

Natural resources are commodities that exist on the planet naturally. Here, commodities that are obtained from these natural resources are tagged as "commodity" only. Some examples from FINTURK datasets are as below:

- For the sentence "Petrol fiyatları çarşamba günü 5.5 yılın en düşük düzeyinden kapanmıştı." (Petrol prices closed at a 5.5-year low on Wednesday.), the word "petrol" is tagged as a natural resource.

- For the sentence "Bir süre sonra parasını altın standardına bağlayıp doları tahtından düşürecek diyorlar." (They say that after a while she will tie her money to the gold standard and dethrone the dollar.), the word "altın" (gold) is tagged as a natural resource.

- For the sentence "Satışlar diğer metallere de sıçradı, kurşun 30 ayın en düşük seviyesine inerken, çinko ve alüminyum son dokuz ayın en düşük düzeylerini test etti." (Sales jumped to other metals as well, with lead hitting 30-month lows, while zinc and aluminum tested nine-month lows.), words "kurşun" (lead), "çinko" (zinc), "alüminyum" (aluminum) are annotated as separate natural resource commodity named entities.

As stated above, commodities that are obtained from natural resources are tagged as "commodity". An example to this is "steel" which is an alloy obtained using mainly iron, carbon and other elements. Since unlike iron element, steel itself is not available in the nature but still a commodity, it is tagged so.

- For the sentence "...Katar, dünyanın bir numaralı sıvılaştırılmış doğalgaz ihracatçısı ve Basra Körfezi'ndeki diğer enerji üreticileri gibi petrol, gaz ihracatı dışında petrokimya, alüminyum ve çelik fabrikalarıyla ekonomisini çeşitlendirmeye çalışıyor." (Qatar, the world's number one exporter of liquefied natural gas and, like other energy producers in the Persian Gulf, is trying to diversify its economy with its petrochemical, aluminum and steel factories, apart from petrol and gas exports.), expressions "doğalgaz", "petrol", "gaz" and "alüminyum" are tagged as separate natural resource named entities. Note that, "gaz" (gas) refers to "natural gas" and it is clear from the context. Moreover, word "çelik" (steel) is tagged as a commodity.

3.2.4.12 Gathering

Gathering is a meeting-like event where people get together. Some examples are, meetings, summits, fairs, exhibitions and so on. In FINTURK, there are both financial and non-financial gatherings. In other words; this category is divided into two sub categories.

Financial Gathering

Gatherings around a domain which is financial or a closely related field. Examples from FINTURK include:

- For the sentence "Ruhani, Başkent Tahran'da düzenlenen 'Milli Ekonomi Konferansı'nda yaptığı konuşmada,..." (Ruhani, in his speech at the "National Economy Conference" held in the capital Tehran, ...), expression "Milli Ekonomi Konferansı" (National Economy Conference) is tagged as a single financial gathering named entity.

- For the sentence "Dünya Ekonomik Forumu (WEF), gelecek hafta yapılacak Davos Zirvesi öncesinde Küresel Riskler 2015 raporunu yayınladı." (The World Economic Forum (WEF) has released its Global Risks 2015 report ahead of next week's Davos Summit.), the expression "Davos Zirvesi" (Davos Summit) is tagged as a financial gathering. Here, the official name of the gathering is not "Davos Zirvesi" however, it is a well-known alias. Actually, this is a financial

event organized by World Economic Forum (WEF) and named as "World Economic Forum" officially as well. In this example sentence “World Economic Forum (WEF)” is tagged as a financial organization since it is an international foundation which has the same name with its event. As seen in this example, annotation is done via the meaning that strongly makes sense to the annotator, out of the context of the text.

Non-Financial Gathering

Non financial gatherings are those which held in fields other than finance or closely related fields to finance. A few examples from FINTURK are as follows:

- For the sentence “87. Pitti Uomo moda fuarına dünyanın pek çok ülkesinden tekstil ve konfeksiyon firması katılıyor”. (Textile and apparel companies from many countries are participating in the 87th Pitti Uomo fashion fair.), expression “87. Pitti Uomo moda fuarı” (87th Pitti Uomo fashion fair) is tagged as a non-financial gathering which is a gathering in fashion and textile domains.

- For the sentence “17-20 Ocak’ta Almanya’nın Hannover kentinde düzenlenen ‘Domotex 2015 Uluslararası Halı ve Zemin Kaplamaları Fuarı’na Güneydoğulu iş adamları yoğun ilgi gösterdi.” (Businessmen from the Southeast showed great interest in the ’Domotex 2015 International Carpet and Floor Coverings Fair’ held in Hannover, Germany on 17-20 January.), expression “Domotex 2015 Uluslararası Halı ve Zemin Kaplamaları Fuarı” (Domotex 2015 International Carpet and Floor Coverings Fair) is tagged as a non-financial gathering which is a fair in the carpet business sector.

3.2.4.13 Tax

This category is for various tax types and also for their abbreviations. Examples include “KDV” (Abbreviation for “value-added-tax”), “ÖTV” (Abbreviation for “excise duty”), “Emlak Vergisi” (Real property tax) etc.

3.2.4.14 Financial Product

These are products which can be gathered from all kinds of corporations in the financial sector. A financial product may be a bank loan or it may be a specific type of insurance policy. In addition, financial assets such as bonds, stocks etc. are also under this category. Examples from FINTURK datasets include: “kredi kartı” (credit card), “konut kredisi” (mortgage loan), “döviz mevduatı” (foreign exchange deposit), “hayat sigortası” (life insurance).

- For “...Fed’in tahvil alım programını sonlandırması euronun önemli ölçüde kan kaybetmesine neden olmuştu.” (The Fed’s termination of its bond-buying program caused the euro to lose significant blood.), the word “tahvil” (bond) is tagged as financial product.

- For “ . . . birikimlerini faizsiz enstrümanlarda değerlendirmek isteyen vatandaşlarımız, BES’e girdiklerinde katkı paylarını kısmen kira sertifikalarında kısment katılım bankalarının katılım hesaplarında ve kısmen de Borsa İstanbul’un katılım endeksindeki firmaların hisse senetlerinde değerlendiriyor.” (Our citizens, who want to invest their savings in interest-free instruments, use their contributions partly in lease certificates, partly in participation accounts of participation banks and partly in stocks of companies in Borsa Istanbul’s participation index.), words “kira sertifikalarında” (in lease certificates), “hisse senetlerinde” (stocks) are tagged as instrument entities. Whereas “katılım hesaplarında” (participation accounts) is tagged as a financial product which is a savings account in a bank which is categorized under financial product.

3.2.4.15 Lottery

In financial newspapers,news related to different games of chance may take place since they are about money somehow. Since FINTURK datasets are originated from news texts from Dünya newspaper, we encountered with some news about games of chance and a type of named entity named “lottery” is formed. Examples of this category from FINTURK datasets are as follows:

- In “...Piyangonun, toto, loto, iddia vb. şans oyunları oynamanın hükmü nedir?...” (What is the ruling on playing lottery, pool, lotto, bets and similar games of

chance?...), words “piyango” (lottery), “toto” (pool), “loto” (lotto), “iddia” (bets) are tagged as different lottery entities.

3.2.5 Inter-Annotator Agreement

FINTURK datasets were annotated entirely by ourselves. Annotating a dataset by a single annotator may lead to some annotation errors and our FINTURK datasets are not 100% error free, annotation was a long and tiring process and it is natural to have mistakes. Apart from these mistakes, annotation made by an annotator may contain subjectivity since as we mentioned in many places in the annotation guidelines that we provided for FINTURK in the earlier section, some annotation decides are closely related to the annotator's understanding from the context.

Since the annotation was done by a single annotator, to see the coherence of the annotation, we constructed 2 small sub datasets from FINTURK500K-PARENTS-BIO and FINTURK500K-CHILDREN-BIO datasets and these small datasets were annotated in BIO format by another annotator who has a long professional experience in finance domain. Then, using the annotations provided by the second annotator and our original annotations for the related tokens (small datasets), we measured the inter-annotator agreement using Cohen's kappa statistic. Cohen's kappa statistic was firstly proposed by Cohen[67] in 1960 is a metric which measures the level of inter-annotator agreement between two annotators/raters in a classification. According to Cohen[67], "The coefficient κ is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration.", in other words, it takes the effect of chance into account when measuring the agreement.

During the Cohen's kappa calculations, we used the "cohen_kappa _score" function [1] under the skicit learn library of Python.

Kappa statistic may take values from -1 to 1 and interpreted as follows[68]:

- $\kappa = 0$: Agreement is equivalent to chance

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen _kappa _score.html

- κ is between 0.10-0.20 : Slight agreement
- κ is between 0.21-0.40 : Fair agreement
- κ is between 0.41-0.60 : Moderate agreement
- κ is between 0.61-0.80 : Substantial agreement
- κ is between 0.81-0.99 : Near perfect agreement
- κ = 1 : Perfect agreement

In addition, in [68] it is stated that: "Negative values indicate that the observed agreement is worse than what would be expected by chance. An alternative interpretation offered is that kappa values below 0.60 indicate a significant level of disagreement."

Moreover, aforementioned 2 sub datasets from FINTURK500K-PARENTS-BIO and FINTURK500K-CHILDREN-BIO datasets, contain randomly selected sentences from main datasets. However, to expand on this "randomness", care was taken to include at least one instance from all types of NEs in these sub datasets and the sentences containing these NEs were selected randomly.

For the sub dataset, which contains randomly selected sentences and totally 2389 tokens (including punctuations as separate tokens) from FINTURK500K-PARENTS-BIO dataset, the Cohen's kappa score was approximately 0.714 and may be interpreted as a substantial agreement between two annotators. Moreover, for the second sub dataset that includes randomly selected sentences and totally 2528 tokens (including punctuations as separate tokens) from FINTURK500K-CHILDREN-BIO dataset, Cohen's kappa score was approximately 0.75 which again can be interpreted as a substantial agreement. It is important to note that, the new annotator has read our annotation guidelines for FINTURK NEs which we presented in the scope of this thesis 3.2.4, we also made some verbal explanations regarding to the parts of the annotation guidelines that are unclear and then we revised our guideline as well. Considering that FINTURK500K-PARENTS-BIO contains less number of NEs than the other dataset,it may be expected that the agreement score for PARENTS version would be higher.However, the PARENTS sub dataset was provided to the new

annotator firstly, and after completing that annotation the annotator started to annotate the CHILDREN sub dataset, this may be a reason that the kappa (κ) score of FINTURK500K-PARENTS-BIO sub dataset is slightly lower than the kappa (κ) score for FINTURK500K-CHILDREN-BIO sub dataset.

CHAPTER 4

EXPERIMENTS AND RESULTS

In the scope of this thesis, experiments with 8 different models were conducted for 4 datasets, namely TUR3, EXTEND7 and FINTURK(PARENTS and CHILDREN versions). In addition, some additional experiments were performed with language-specific BERT models. In the evaluation step, the primary evaluation metrics of CoNLL were used.

Furhermore, experiments were conducted in Google Colab environment using Google's GPU's that are provided to users for a limited time.

Moreover, models that are used in these experiments are available in Hugging Face repository which embodies most of the transformer based language models that are implemented with PyTorch.On their website [1] there are also many pretrained models for many languages or various domains which are provided by different people or groups. In all subsections related to experiments, details of models including the Hugging Face platform addresses are provided.

4.1 Information About All Datasets Used

4.1.1 FINTURK Dataset

As explained in 3.2.1, FINTURK dataset is constructed by manually annotating news text from Dünya newspaper. Dataset was splitted into 3 parts as train, development and test splits. Throughout the experiments, 3 different split versions of same datasets were experimented with all of the models. In the first split version, 60% of the data

[1] https://huggingface.co/

is reserved for training, 20%for development and 20% for testing. Moreover, for the second and third split versions, 80% of the data was used during training the models and 10% for development and 10% for testing. Token distributions of these splits are as seen in Table 4.1. Furthermore, the only difference between second and third split is that; the development data of split 2 is the testing data for split v3 and vice versa. Since in our newly proposed datasets we have many NEs, the test data file of split v2 did not contain "LOTTERY" NE instance and so, we replaced the development and test files and this brought us to split v3. It is important to note that, as it can be realized from the tables 4.4 and 4.5, "FIN-GATHERING" which is among the CHILDREN tags, is not represented with any instance in the test file of split v3. By constructing such different splits from the same datasets, we obtained the chance of observing the model performances for all of the distinct NEs.

Moreover, distributions of different named entity tags throughout different partitions of FINTURK PARENTS Dataset versions are shown in Table 4.2 and 4.3 and similar table is available in tables 4.4 and 4.5 for FINTURK CHILDREN Dataset versions.

Table 4.1: Token Distributions of FINTURK500K, TUR3 and EXTEND7 Datasets

Partitions	**FINTURK500K SPLIT v1 (60-20-20)**	**FINTURK500K SPLIT v2 (80-10-10)**	**FINTURK500K SPLIT v3 (80-10-10)**	**TUR3 (80-10-10)**	**EXTEND7 (80-10-10)**
Train	301,753	402,203	402,203	437,465	408,184
Development	100,450	50,220	50,234	33,818	36,661
Test	100,454	50,234	50,220	47,249	24,710
TOTAL	502,657	502,657	502,657	518,532	469,555

Table 4.2: Number of Named Entity Tags in FINTURK500K PARENTS Datasets' Different Partitions Annotated with BIO Schema

Dataset Split Version	**PARENTS BIO SPLIT v1 (60-20-20)**			**PARENTS BIO SPLIT v2 (80-10-10)**			**PARENTS BIO SPLIT v3 (80-10-10)**		
Named Entity Tag	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
B-MONETARY-VALUE	2286	788	741	3074	338	403	3074	403	338
I-MONETARY-VALUE	2503	726	786	3229	356	430	3229	430	356
B-CURRENCY	2549	934	932	3483	443	489	3483	489	443
I-CURRENCY	55	28	35	83	20	15	83	15	20
B-LOCATION	7597	2563	2595	10160	1205	1390	10160	1390	1205
I-LOCATION	962	391	335	1353	163	172	1353	172	163
B-PERCENT	3104	869	1163	3973	694	469	3973	469	694
I-PERCENT	3192	859	1190	4051	707	483	4051	483	707
B-PERSON	3418	1254	1148	4672	558	590	4672	590	558
I-PERSON	1691	678	529	2369	246	283	2369	283	246
B-COMMODITY	1971	649	407	2620	211	196	2620	196	211
I-COMMODITY	338	119	66	457	29	37	457	37	29
B-ORG	5429	2010	1961	7439	973	988	7439	988	973
I-ORG	6891	2383	2206	9274	1042	1164	9274	1164	1042
B-DATE	2304	686	763	2990	396	367	2990	367	396
I-DATE	1923	608	700	2531	381	319	2531	319	381
B-INDEX	569	226	184	795	61	123	795	123	61
I-INDEX	727	472	244	1199	117	127	1199	127	117
B-BUSINESS-SECTOR	3107	1092	832	4199	450	382	4199	382	450
I-BUSINESS-SECTOR	1661	684	319	2345	211	108	2345	108	211
B-TAX	97	75	36	172	15	21	172	21	15
I-TAX	55	54	31	109	12	19	109	19	12
B-GATHERING	240	55	91	295	39	52	295	52	39
I-GATHERING	715	180	140	895	73	67	895	67	73
B-FIN-PROD	675	161	211	836	124	87	836	87	124
I-FIN-PROD	444	106	86	550	50	36	550	36	50
B-LOTTERY	15	4	7	19	7	N/A	19	N/A	7
I-LOTTERY	1	4	6	5	6	N/A	5	N/A	6
B-TIME	37	8	5	45	2	3	45	3	2
I-TIME	83	20	5	103	3	2	103	2	3
O	247114	81764	82700	328878	41288	41412	328878	41412	41288

Table 4.3: Number of Named Entity Tags in FINTURK500K PARENTS Datasets' Different Partitions Annotated with NON-BIO Schema

Dataset Split Version	**PARENTS NON-BIO SPLIT v1 (60-20-20)**			**PARENTSNON-BIO SPLIT v2 (80-10-10)**			**PARENTS NON-BIOSPLIT v3 (80-10-10)**		
Named Entity Tag	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
MONETARY-VALUE	4789	1514	1527	6303	694	833	6303	873	694
CURRENCY	2604	962	967	3566	463	504	3566	504	463
LOCATION	8559	2954	2930	11513	1368	1562	11513	1562	1368
PERCENT	6296	1728	2353	8024	1401	952	8024	952	1401
PERSON	5109	1932	1677	7041	804	873	7041	873	804
COMMODITY	2309	768	473	3077	240	233	3077	233	240
ORG	12320	4393	4167	16713	2015	2152	16713	2152	2015
DATE	4227	1294	1463	5521	777	686	5521	686	777
INDEX	1296	698	428	1994	178	250	1994	250	178
BUSINESS-SECTOR	4768	1776	1151	6544	661	490	6544	490	661
TAX	152	129	67	281	27	40	281	40	27
GATHERING	955	235	231	1190	112	119	1190	119	112
FIN-PROD	1119	267	297	1386	174	123	1386	123	174
LOTTERY	16	8	13	24	13	N/A	24	N/A	13
TIME	120	28	10	148	5	5	148	5	5
O	247114	81764	82700	328878	41288	41412	328878	41412	41288

Table 4.4: Number of Named Entity Tags in FINTURK500K CHILDREN Datasets' Different Partitions Annotated with BIO Annotation Schema

*

Dataset Split Version	**CHILDREN BIO SPLIT v1 (60-20-20)**			**CHILDREN BIO SPLIT v2 (80-10-10)**			**CHILDREN BIO SPLIT v3 (80-10-10)**		
Named Entity Tag	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
B-MONETARY-VALUE	2286	788	741	3074	338	403	3074	403	338
I-MONETARY-VALUE	2503	726	786	3229	356	430	3229	430	356
B-CURRENCY	2549	934	932	3483	443	489	3483	489	443
I-CURRENCY	55	28	35	83	20	15	83	15	20
B-LOCATION	7597	2563	2595	10160	1205	1390	10160	1390	1205
I-LOCATION	962	391	335	1353	163	172	1353	172	163
B-PERCENT	3104	869	1163	3973	694	469	3973	469	694
I-PERCENT	3192	859	1190	4051	707	483	4051	483	707
B-PERSON	3418	1254	1148	4672	558	590	4672	590	558
I-PERSON	1691	678	529	2369	246	283	2369	283	246
B-NAT-RES	906	258	219	1164	105	114	1164	114	105

Table 4.4: Continued

I-NAT-RES	101	18	11	119	2	9	119	9	2
B-NON-FIN-ORG	3533	1246	1096	4779	537	559	4779	559	537
I-NON-FIN-ORG	4731	1581	1373	6312	659	714	6312	714	659
B-DATE	2304	686	763	2990	396	367	2990	367	396
I-DATE	1923	608	700	2531	381	319	2531	319	381
B-INDEX	569	226	184	795	61	123	795	123	61
I-INDEX	727	472	244	1199	117	127	1199	127	117
B-FIN-ORG	1896	764	865	2660	436	429	2660	429	436
I-FIN-ORG	2160	802	833	2962	383	450	2962	450	383
B-BUSINESS-SECTOR	3107	1092	832	4199	450	382	4199	382	450
I-BUSINESS-SECTOR	1661	684	319	2345	211	108	2345	108	211
B-AGRIC-COMMODITY	709	324	119	1033	88	31	1033	31	88
I-AGRIC-COMMODITY	166	91	33	257	27	6	257	6	27
B-TAX	97	75	36	172	15	21	172	21	15

Table 4.4: Continued

I-TAX	55	54	31	109	12	19	109	19	12
B-NON-FIN-GATHERING	184	45	87	229	39	48	229	48	39
I-NON-FIN-GATHERING	549	132	131	681	73	58	681	58	73
B-FIN-PROD	675	161	211	836	124	87	836	87	124
I-FIN-PROD	444	106	86	550	50	36	550	36	50
B-COMMODITY	356	67	69	423	18	51	423	51	18
I-COMMODITY	71	10	22	81	N/A	22	81	22	N/A
B-FIN-GATHERING	56	10	4	66	N/A	4	66	4	N/A
I-FIN-GATHERING	166	48	9	214	N/A	9	214	9	N/A
B-LOTTERY	15	4	7	19	7	N/A	19	N/A	7
I-LOTTERY	1	4	6	5	6	N/A	5	N/A	6
B-TIME	37	8	5	45	2	3	45	3	2
I-TIME	83	20	5	103	3	2	103	2	3
O	247114	81764	82700	328878	41288	41412	328878	41412	41288

Table 4.5: Number of Named Entity Tags in FINTURK500K CHILDREN Datasets' Different Partitions Annotated with NON-BIO Annotation Schema

Dataset Split Version	**CHILDREN NON-BIO SPLIT v1 (60-20-20)**			**CHILDREN NON-BIO SPLIT v2 (80-10-10)**			**CHILDREN NON-BIO SPLIT v3 (80-10-10)**		
Named Entity Tag	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
MONETARY-VALUE	4789	1514	1527	6303	694	833	6303	833	694
CURRENCY	2604	962	967	3566	463	504	3566	504	463
LOCATION	8559	2954	2930	11513	1368	1562	11513	1562	1368
PERCENT	6296	1728	2353	8024	1401	952	8024	952	1401
PERSON	5109	1932	1677	7041	804	873	7041	873	804
NAT-RES	1007	276	230	1283	107	123	1283	123	107
NON-FIN-ORG	8264	2827	2469	11091	1196	1273	11091	1273	1196
DATE	4227	1294	1463	5521	777	686	5521	686	777
INDEX	1296	698	428	1994	178	250	1994	250	178
FIN-ORG	4056	1566	1698	5622	819	879	5622	879	819
BUSINESS-SECTOR	4768	1776	1151	6544	661	490	6544	490	661
AGRIC-COMMODITY	875	415	152	1290	115	37	1290	37	115
TAX	152	129	67	281	27	40	281	40	27
NON-FIN-GATHERING	733	177	218	910	112	106	910	106	112
FIN-PROD	1119	267	297	1386	174	123	1386	123	174
COMMODITY	427	77	91	504	18	73	504	73	18
FIN-GATHERING	222	58	13	280	N/A	13	280	13	N/A
LOTTERY	16	8	13	24	13	N/A	24	N/A	13
TIME	120	28	10	148	5	5	148	5	5
O	247114	81764	82700	328878	41288	41412	328878	41412	41288

4.1.2 TUR3 Dataset

TUR3 dataset is a news dataset from a Turkish newspaper Milliyet, covers articles from January 1 1997 through September 12 1998 [11]. This dataset was proposed by Tür et al. in 2003[11]. The number of tokens contained in train, development and test portions of TUR3 dataset are as seen in Table 4.1. TUR3 dataset's train, development and test split ratios are %80,%20 and %20 respectively.

TUR3 dataset contains 3 different types of named entities, namely "PERSON", "ORGANIZATION" and "LOCATION". Distribution of these named entities between train, development and test splits for TUR3 dataset is available in Table 4.6. In the table, both NE counts and token-tag counts are provided.

4.1.3 EXTEND7 Dataset

EXTEND7 dataset is a different version of TUR3 dataset[11] where annotation is extended by Şeker and Eryiğit [39] as containing 7 different named entities other than "O"(other). This dataset contains "PERSON", "ORGANIZATION", "LOCATION", "DATE", "TIME", "MONEY" and "PERCENT" entities.

Since in our paper [45] we mentioned these datasets as TUR3 and EXTEND7, here in this thesis same naming convention is used.

Similar to TUR3 dataset, EXTEND7 dataset's train, development and test split ratios are %80,%20 and %20 respectively. Moreover, named entity distribution in train, development and test splits of EXTEND7 is presented in Table 4.6.

TUR3 and EXTEND7 datasets are formerly annotated with RAW tags where there is no prefix such as "B-" or "I-" that indicate the boundaries of the named entity.This schema of tagging, i.e. using RAW tags (such as "PERSON"), actually can be easily converted to "Inside-Outside notation(IO)" where the tag is preceded with a "I-" such as "I-PERSON". Indig et al. mentions that [9] IO notation and the prefixless notation are the two inferior representations for tags where the tag consists only of the chunk type. They also mention that these variants cannot distinguish between subsequent chunks of the same type, but they are easy to use for seaching nonconsecutive chunks

Table 4.6: Number of Named Entity Tags in TUR3 and EXTEND7 Datasets Partitions Annotated with NON-BIO Schema

Dataset Split Version	**TUR3 SPLITS (80-10-10) NE count/ Tag count**			**EXTEND7 SPLITS (80-10-10) NE count/ Tag count**		
Named Entity	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
PERSON	13542/19983	1055/1563	1603/2384	13540/19957	1056/1552	658/1019
LOCATION	8913/10206	779/903	1125/1331	8915/10218	779/903	637/732
ORGANIZATION	8370/13947	768/1248	871/1688	8347/13993	769/1251	408/822
DATE	-	-	-	1213/2165	150/282	116/259
TIME	-	-	-	125/266	23/27	19/32
MONEY	-	-	-	540/1654	51/178	39/127
PERCENT	-	-	-	560/1156	88/174	62/124

due to their simplicity. In the scope of our thesis, we did not convert these datasets to another format, however, in the performance evaluation phase, we converted the annotation format to BIO schema using a tool mentioned in detail under the sub chapter 4.2.

4.2 Evaluation

In the literature, there are 2 main evaluation metrics definitions that are used extensively.One widely used evaluation metrics definition is by CoNLL [15]. According to this definition, precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if it is an exact match of the corresponding entity in the data file. F-measure is calculated by the equation: $F - measure = \frac{2*Precision*Recall}{Precision+Recall}$

Other commonly used evaluation metric set is presented by MUC. According to this approach, metrics adopted more loose matching conditions which allow for partial credit when partial span or wrong type detection happened. Unlike the CoNLL eval-

uation, where the system's predictions should match with the boundaries and types of named entities exactly for a full credit, in MUC evaluation there is more chance to get credit. If the system detects the entity type correctly but fails to detect the exact boundaries of that named entity, then the score is not zero. The converse is also true, if the system detects the correct boundary for a named entity but fails to detect the correct type, then again partial credit is received.[69]

In the scope of this thesis, CoNLL evaluation metric is adopted since it gives more strict accuracy for the evaluation of the performance.

When considering the entity-wise evaluation (as in CoNLL evaluation) rather than the token-wise evaluation, it can be claimed that computing the confusion matrix is not straightforward. For example, in Table 4.7 an example for a true positive instance is shown where the predicted and ground truth labels match for the entire boundary of the named entity. Another straightforward example is shown in Table 4.8 where a false negative instance occurs. However, when the model makes mistake in the prediction of the boundary of a named entity, as seen in Table 4.9, that mistake is counted as false negative and false positive at the same time.

Table 4.7: An Example for True Positive

Token	**Ground Truth Label**	**Predicted Label**
ODTÜ	B-ORG	B-ORG
Mühendislik	I-ORG	I-ORG
Fakültesi	I-ORG	I-ORG
sıralamada	O	O
birinci	O	O
oldu	O	O

The confusion matrix for the instances presented in tables 4.7, 4.8 and 4.9is as shown in Table 4.10. As seen from the table, for the three examples above, we have 1 true positive, 1 false positive and 2 false negatives. 1 false positive and 1 false negative comes from the instance on Table 4.9 where the boundary is predicted erroneously. Here, we both see a predicted entity when there is not a such NE (for "Afrika" token) and this causes a false positive error. On the other hand, we see that prediction of the

Table 4.8: An Example for False Negative

Token	Ground Truth Label	Predicted Label
Konut	B-FIN-PROD	O
kredisi	I-FIN-PROD	O
faizlerinde	O	O
düşüş	O	O
yaşandı	O	O

Table 4.9: An Example for Double Error

Token	Ground Truth Label	Predicted Label
Sahra	B-LOCATION	O
Altı	I-LOCATION	O
Afrika	I-LOCATION	B-LOCATION
göç	O	O
veriyor	O	O

labels for NE "Sahra Altı Afrika" is not correct and this gives rise to a false negative error. At the end, we conclude with 2 different errors originated from a single wrong prediction.

During the labelwise evaluation process of experiments with BIO tagged data, a Perl script [2] that was said to be written by Sabine Buchholz from Tilburg University, was used. Furhermore, for the evaluation of the labelwise performances related to NON-BIO tagged datasets,the predicted dataset (i.e. predicted version of the test file) is retrieved and the labels in NON-BIO format, namely in the raw format (such as "PERSON") were converted to the BIO(IOB2) annotation schema and the same Perl script that was used formerly for the BIO-tagged datasets were used to compute labelwise performance scores. Label conversion was done via using an IOB format converter and corrector[3] developed by the Research Group of Language Technology, NYTK. According to the README file on the groups github page, "emIOBUtils is

[2] https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt
[3] https://github.com/nytud/emIOBUtils

Table 4.10: Confusion Matrix for the Instances Above

		Predicted	
		Negative	**Positive**
Ground Truth	**Negative**	**True Negative TN=0**	**False Positive TP=1**
	Positive	**False Negative FN=2**	**True Positive TP=1**

the Python rewrite of CoreNLP's IOBUtils which is written in JAVA. It can take any (possibly ill-formed) IOB span input and convert/correct it according to the specified output style." and the supported formats are: IOB, IOE, BIO/IOB, IO, SBIEO/IOBES and NOPREFIX. Here, "NOPREFIX" format corresponds to raw/NON-BIO format in our thesis.

By converting the predicted test datasets of NON-BIO tagged datasets to BIO format, we aimed to compare the effect of training the same dataset with different annotation schemas, on the performance of the models. We trained same FINTURK datasets with both BIO and NON-BIO annotation using same models and same setting and parameters. Moreover, in the end, with the help of this conversion, we had the opportunity to compare different prediction datasets with different annotation formats by using the same CoNLL evaluation tool that was used extensively throughout the researches in this field.

Speaking of evaluation, an interesting study by Alshammari et al.[70] is worth to mention.Authors studied seven annotation schemes (IO, IOB, IOE, IOBES, BI, IE and BIES) and their impact on NER task with five different classifiers. They used a dataset consisting of more than 62,000 tokens which were parts of 27 medical Arabic articles. They observed that IO annotation schema showed the best performance with an F-score of 84.44%. Second best performed annotation schema was BIES schema with F1-score as 72.78%. In the scope of this thesis BIO (same as the IOB format mentioned in this paper) annotation schema was used as mentioned before. In the article, researchers observed an F1-score of 63.18% for IOB annotation schema which performed the third worst among all schemas. It is worth to recall that, IO annota-

tion schema that performed best is very close to the raw annotation schema, actually adding "I-" prefix in front of all tags other than "O" (other) tag will convert raw tags to IO tags. Thus, from this study we can also deduce that raw annotation schema may perform better than BIO schema in our case.

4.3 Experiments with BERT-Base Multilingual Cased Model

As mentioned in 3,BERT-Base Multilingual Cased model was firstly relesased via a paper by Devlin et al. [2] in 2018. It is a pretrained model via using a Wikipedia corpus from 104 different languages. Since the maximum sequence length that BERT supports is 512, during experiments the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. The model in Hugging Face library[71] was used throughout the experiments for BERT-Base Multilingual Cased model. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[4]. Similarly, other training arguments/parameters are set to default.

Results of experiments with BERT-Base Multilingual Cased model on evaluation and test dataset splits of FINTURK500K BIO and FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are shown in tables 4.11 and 4.12 respectively.

As it can be seen from the both tables, we used the 60% of the datasets for training for the first split version of FINTURK PARENTS and CHILDREN datasets. The size of training set affects the performance,namely the F-score approximately 1% percent.

Comparing the scores gathered with BIO and NON-BIO annotated versions of same datasets; it can be observed from the tables that, the performance of the same model on NON-BIO dataset versions are slightly higher than the performance on BIO datasets for PARENTS versions of FINTURK dataset splits. However, for CHILDREN versions of FINTURK dataset splits, we can observe approximately equal F1-scores with BIO and NON-BIO formats. Thus, for these experiments, we can conclude that there is not a meaningful difference in results considering the annotation format differences.

[4] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

Table 4.11: Precision, Recall and F1-Scores of BERT-Base Multilingual Cased Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	**FINTURK500K-PARENTS-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v3 (80-10-10)**
Precision	89.17	89.98	89.20	88.41	90.04	87.83
Recall	90.35	91.53	91.21	89.63	91.27	89.90
F1-Score	89.75	90.75	90.19	89.02	90.65	88.86

Table 4.12: Precision, Recall and F1-Scores of BERT-Base Multilingual Cased Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	**FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)**	**TUR3 (80-10-10)**	**EXTEND7 (80-10-10)**
Precision	89.53	90.64	89.93	87.97	90.17	88.05	93.29	91.15
Recall	90.45	91.29	91.20	89.62	91.15	89.70	93.41	92.05
F1-Score	89.99	90.96	90.56	88.79	90.66	88.87	93.35	91.60

Comparing the different datasets, it is observed that the best performance with BERT multilingual model is on TUR3 dataset.

Furthermore, considering the PARENTS and CHILDREN versions of the FINTURK dataset, a slightly larger F1-score is observed with PARENTS dataset. This difference makes sense because of the fact that the PARENTS dataset contains less number of labels and this leads the model to learn the tagging easier.

In Table 4.13,labelwise performances of BERT-Base Multilingual Cased model with FINTURK500K-PARENTS-BIO dataset's different split versions is shown. Among the labelwise performances, NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY -VALUE" are common in lists of top 5 highest F1-scores in all 3 different splits of the same dataset. Moreover, NEs "FIN-PROD", "BUSINESS-SECTOR" and "GATHERING" are common NEs in the lists of top 5 lowest F1-scores in all 3 different splits.

In Table 4.14,labelwise performances of BERT-Base Multilingual Cased model with FINTURK500K-PARENTS-NON-BIO dataset's different split versions is shown. Among the labelwise performances, NEs "PERSON", "PERCENT", "MONETARY-VALUE" and "CURRENCY" are common in lists of top 5 highest F1-scores in all 3 different splits of the same dataset. Moreover, NEs "BUSINESS-SECTOR", "FIN-PROD" and "LOTTERY" are common NEs in the lists of top 5 lowest F1-scores in all different splits. Here, it is important to recall that, "LOTTERY" NE does not exist in the test split of the second version of splits, so we can compare the entity-wise scores of this NE for just first and third split versions.

Similar to the PARENTS BIO and NON-BIO datasets of FINTURK, we can see the labelwise performances of BERT-Base Multilingual Cased model using different split versions of FINTURK500K-CHILDREN-BIO and FINTURK500K-CHILDREN-NON-BIO datasets are shown in tables 4.15 and 4.16, respectively.

Table 4.13: Scores of BERT-Base Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	58.56	64.27	52.86	74.04	81.94	69.78	65.39	72.04	60.15
COMMODITY	82.65	87.50	87.68	79.61	82.14	87.68	81.10	84.74	87.68
CURRENCY	98.60	98.35	98.86	97.86	97.34	97.97	98.22	97.84	98.42
DATE	90.14	92.41	90.63	91.33	92.92	90.86	90.73	92.66	90.75
FIN-PROD	69.88	59.15	75.00	54.98	48.28	65.32	61.54	53.16	69.83
GATHERING	63.04	71.05	68.29	63.74	51.92	71.79	63.39	60.00	70.00
INDEX	89.89	89.06	88.89	91.85	92.68	91.80	90.86	90.84	90.32
LOCATION	92.11	91.39	92.85	91.80	93.10	92.61	91.95	92.24	92.73
LOTTERY	0.00	N/A	100.00	0.00	N/A	14.29	0.00	N/A	25.00
MONETARY-VALUE	97.33	97.69	97.27	93.52	94.54	94.97	95.39	96.09	96.11
ORG	88.29	88.62	89.89	89.60	91.40	91.37	88.94	89.99	90.62
PERCENT	98.10	96.78	98.98	97.34	95.96	98.27	97.71	96.37	98.63
PERSON	96.96	97.42	98.04	97.30	95.77	98.39	97.13	96.59	98.21
TAX	82.86	90.48	92.86	80.56	90.48	86.67	81.69	90.48	89.66
TIME	83.33	75.00	100.00	100.00	100.00	100.00	90.91	85.71	100.00

Table 4.14: Scores of BERT-Base Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	58.90	68.51	53.70	72.44	80.89	69.49	64.98	74.19	60.58
COMMODITY	85.38	77.36	89.80	80.34	83.67	83.41	82.78	80.39	86.49
CURRENCY	98.91	98.75	98.62	97.95	97.53	97.73	98.42	98.13	98.17
DATE	91.44	94.57	90.10	91.20	94.82	90.10	91.32	94.69	90.10
FIN-PROD	69.41	60.61	72.82	56.46	47.06	60.48	62.27	52.98	66.08
GATHERING	65.93	66.67	69.77	65.93	53.85	76.92	65.93	59.57	73.17
INDEX	89.89	89.84	90.32	91.85	93.50	91.80	90.86	91.63	91.06
LOCATION	91.79	92.64	93.74	92.89	92.64	93.43	92.34	92.64	93.59
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	97.86	98.16	98.78	93.21	93.02	96.72	95.48	95.52	97.74
ORG	88.88	89.29	91.91	89.52	90.47	91.34	89.20	89.87	91.62
PERCENT	98.26	98.06	98.70	97.16	96.60	98.27	97.71	97.32	98.48
PERSON	96.70	95.59	97.70	96.87	95.26	98.92	96.78	95.42	98.31
TAX	80.00	90.91	86.67	80.00	95.24	92.86	80.00	93.02	89.66
TIME	80.00	66.67	100.00	80.00	66.67	100.00	80.00	66.67	100.00

Table 4.15: Scores of BERT-Base Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	70.99	56.41	77.91	78.15	70.97	76.14	74.40	62.86	77.01
BUSINESS-SECTOR	59.62	66.32	52.29	71.88	82.98	68.44	65.18	73.72	59.29
COMMODITY	64.13	81.48	60.00	85.51	86.27	83.33	73.29	83.81	69.77
CURRENCY	98.59	98.76	98.87	97.32	97.55	98.65	97.95	98.15	98.76
DATE	91.66	94.26	91.92	92.38	94.01	92.39	92.02	94.13	92.15
FIN-GATHERING	22.22	100.00	N/A	50.00	50.00	N/A	30.77	66.67	N/A
FIN-ORG	90.40	88.94	85.62	89.35	92.06	88.76	89.87	90.47	87.16
FIN-PROD	65.27	56.34	74.31	51.66	45.98	65.32	57.67	50.63	69.53
INDEX	90.37	93.55	88.71	91.85	94.31	90.16	91.11	93.93	89.43
LOCATION	92.38	91.43	93.24	92.49	92.81	92.78	92.44	92.12	93.01
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	97.48	97.94	97.56	94.06	94.54	94.67	95.74	96.21	96.10
NAT-RES	90.82	91.59	92.78	85.84	85.96	85.71	88.26	88.69	89.11
NON-FIN-GATHERING	54.55	75.56	70.27	48.28	70.83	66.67	51.22	73.12	68.42
NON-FIN-ORG	78.87	85.71	80.26	82.45	83.42	81.45	80.62	84.55	80.85
PERCENT	98.10	97.42	98.57	97.77	96.38	98.99	97.93	96.90	98.78
PERSON	96.27	97.29	97.87	96.61	97.12	98.75	96.44	97.21	98.31
TAX	86.49	82.61	92.86	88.89	90.48	86.67	87.67	86.36	89.66
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.16: Scores of BERT-Base Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	62.07	64.86	87.01	75.63	77.42	76.14	68.18	70.59	81.21
BUSINESS-SECTOR	56.70	66.32	50.94	72.32	82.46	66.59	63.56	73.51	57.72
COMMODITY	72.29	77.78	66.67	86.96	82.35	88.89	78.95	80.00	76.19
CURRENCY	98.48	98.95	99.54	98.05	97.11	98.18	98.27	98.02	98.86
DATE	91.35	93.21	90.13	91.59	93.46	90.36	91.47	93.33	90.24
FIN-GATHERING	33.33	66.67	N/A	50.00	50.00	N/A	40.00	57.14	N/A
FIN-ORG	88.67	91.32	87.07	88.98	93.68	88.28	88.82	92.49	87.67
FIN-PROD	71.70	60.29	78.00	54.55	48.24	62.90	61.96	53.59	69.64
INDEX	88.36	92.25	90.32	90.76	96.75	91.80	89.54	94.44	91.06
LOCATION	92.50	91.93	93.50	92.43	93.00	93.27	92.46	92.46	93.38
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.31	98.44	98.78	94.84	94.26	96.42	96.54	96.31	97.58
NAT-RES	88.06	89.81	89.47	80.82	85.09	80.95	84.29	87.39	85.00
NON-FIN-GATHERING	69.32	69.57	72.34	70.11	66.67	87.18	69.71	68.09	79.07
NON-FIN-ORG	78.32	84.03	80.11	81.60	82.68	80.86	79.93	83.35	80.48
PERCENT	98.17	97.21	98.56	96.99	96.38	98.41	97.58	96.79	98.49
PERSON	96.95	97.75	98.05	96.87	95.77	98.92	96.91	96.75	98.48
TAX	86.11	95.45	92.86	88.57	100.00	92.86	87.32	97.67	92.86
TIME	60.00	66.67	100.00	60.00	66.67	100.00	60.00	66.67	100.00

Table 4.17: Scores of BERT-Base Multilingual Cased Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	93.40	88.99	93.51	92.10	93.45	90.52
ORGANIZATION	91.11	90.50	91.73	90.50	91.42	90.50
LOCATION	94.83	96.50	94.58	95.58	94.70	96.04
DATE	-	75.21	-	75.86	-	75.54
TIME	-	89.47	-	89.47	-	89.47
MONEY	-	97.50	-	100.00	-	98.73
PERCENT	-	91.38	-	91.38	-	91.38

From Table 4.15, it can be seen that, NEs "CURRENCY","PERCENT","PERSON"and "MONETARY-VALUE" are common NEs in the lists of top 5 highest F1-scores in all split versions. Additionally, "LOTTERY", "NON-FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" are NEs which are common in lists of lowest labelwise F1-scores for different split versions.

Moreover, as it can be seen from the Table 4.16, NEs "CURRENCY","PERCENT", "PERSON" and "MONETARY-VALUE" are common in lists of top 5 labelwise F1-scores for all 3 split versions of the same dataset, namely, FINTURK500K-CHILDREN-NON-BIO dataset. Again, if we compare the lowest labelwise F1-scores for this model and dataset, we can see that "LOTTERY" and "FIN-PROD" are common in the bottom 5 F1-scores lists for different splits. Again for "LOTTERY" NE, we can just take split version 13 into account since "LOTTERY" tag is not included in the test set of split version 2.

In Table 4.17 labelwise scores of BERT-Base Multilingual Cased model for TUR3 and EXTEND7 datasets are shown. For TUR3 dataset, the highest labelwise F1-score is associated with "LOCATION" NE which is associated with the second highest labelwise score for EXTEND7 dataset. For EXTEND7 dataset, the highest score is for "MONEY". Considering the lowest scores, we see the lowest score for TUR3 dataset is associated with "ORG" which is the third lowest score for the same label in EXTEND7 dataset. Moreover,considering EXTEND7 alone, the lowest F1-score is for "DATE" NE.

4.4 Experiments with Bert Base Turkish Cased (BERTurk) Model

After the success of BERT model, Turkish-specific versions of the model were presented which were trained with different datasets sizes[72]. In this section, we will use BERT-Base Turkish Cased model [21] which is community-driven Turkish-specific BERT model where a vocabulary size of 128k was used througout the training process. In the model's readme file, researchers state that for pretraining and evaluation some datasets which are contributed from the awesome Turkish NLP community, were used. Model is available in Hugging Face library. Since the maximum sequence length that BERT supports is 512, during experiments the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[5]. Similarly, other training arguments/parameters are set to default.

Results of experiments with BERTurk model on evaluation and test dataset splits of FINTURK500K BIO,FINTURK500K NON-BIO, TUR3 and EXTEND7 datasts are as shown in Tables 4.11 and 4.19 respectively.

60% of the datasets were used for training for the first split version and 80% used for training for the second and third split versions of FINTURK PARENTS and CHILDREN datasets. Comparing results for split 12 it may be concluded that the since the size of the training has an approximately 1% impact on the F1-score, however there is no such an affect when comparing split 1&3. Thus, for this model and setting,it may be inferred that increasing the training set size does not have an impact on the performance.

Comparing the BIO and NON-BIO annotated versions of same datasets, there seems to be no substantial difference in performances.

[5] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

Table 4.18: Precision, Recall and F1-Scores of BERT-Turkish Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	FINTURK500K-PARENTS-BIO SPLIT v1(60-20-20)	FINTURK500K-PARENTS-BIO SPLIT v2(80-10-10)	FINTURK500K-PARENTS-BIO SPLIT v3(80-10-10)	FINTURK500K-CHILDREN-BIO SPLIT v1(60-20-20)	FINTURK500K-CHILDREN-BIO SPLIT v2(80-10-10)	FINTURK500K-CHILDREN-BIO SPLIT v3(80-10-10)
Precision	91.97	93.22	91.59	90.85	92.21	89.89
Recall	92.86	93.37	92.91	92.21	93.17	91.94
F1-Score	92.42	93.29	92.25	91.53	92.69	90.90

Table 4.19: Precision, Recall and F1-Scores of BERT-Turkish Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K-PARENTS-NON-BIO SPLIT v2 (80-10-10)	FINTURK500K-PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
Precision	91.75	92.78	90.67	90.67	92.42	90.04	95.95	93.55
Recall	92.89	93.62	92.64	92.35	93.10	91.96	96.19	93.60
F1-Score	92.32	93.20	91.64	91.50	92.76	90.99	96.07	93.58

Comparing the different datasets, it can be concluded that BERT Base Turkish Cased model exhibits the best performance with using TUR3 dataset which is followed by EXTEND7 and FINTURK500K-PARENTS-BIO dataset's second split version.

Moreover, comparing the difference in performance using PARENTS and CHILDREN versions of FINTURK dataset, it can be concluded that for the PARENTS versions of all dataset splits of FINTURK, the performance of the model is approximately 1% higher compared to the corresponding splits of CHILDREN version. This result is understandable since the CHILDREN version contains more labels for the model to learn.

In Table 4.20,labelwise performances of BERT Base Turkish Cased model using different split version of FINTURK500K-PARENTS-BIO dataset different split versions is shown. Among the labelwise performances, NEs "PERSON", "PERCENT"and "MONETARY-VALUE" are common in lists of top 5 highest F1-scores in all 3 different splits of the same dataset.Moreover, NEs "FIN-PROD", "BUSINESS-SECTOR", "GATHERING" are common NEs in the lists of top 5 lowest F1-scores in all 3 different splits.

In Table 4.21,labelwise performances of BERT Base Turkish Cased model with FINTURK 500K-PARENTS-NON-BIO dataset's different split versions is shown. Among the labelwise performances, NEs "PERSON", "PERCENT", "MONETARY-VALUE" and "CURRENCY" are common in lists of top 5 highest F1-scores in all 3 different splits of the same dataset.Moreover, NEs "BUSINESS-SECTOR", "FIN-PROD" and "LOTTERY" are common NEs in the lists of top 5 lowest F1-scores in different splits.

Furthermore, labelwise performance results of BERT Base Turkish Cased Model using different split versions of FINTURK500K-CHILDREN-BIO and FINTURK500K -CHILDREN-NON-BIO datasets are shown in tables; 4.22 and 4.23 respectively.

Table 4.20: Scores of BERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	65.02	72.44	59.93	75.96	83.25	71.78	70.07	77.47	65.32
COMMODITY	85.08	87.50	86.82	89.68	85.71	90.52	87.32	86.60	88.63
CURRENCY	99.03	98.98	99.55	98.82	98.77	99.32	98.93	98.87	99.44
DATE	92.92	94.05	91.67	92.92	94.82	91.67	92.92	94.44	91.67
FIN-PROD	76.65	69.84	75.68	60.66	50.57	67.74	67.72	58.67	71.49
GATHERING	71.03	85.71	70.83	83.52	80.77	87.18	76.77	83.17	78.16
INDEX	91.35	95.12	90.32	91.85	95.12	91.80	91.60	95.12	91.06
LOCATION	94.42	94.66	95.56	94.53	94.39	94.53	94.48	94.53	95.04
LOTTERY	0.00	N/A	100.00	0.00	N/A	85.71	0.00	N/A	92.31
MONETARY-VALUE	98.33	98.68	98.19	95.28	93.05	96.15	96.78	95.79	97.16
ORG	92.57	92.81	93.33	92.71	94.03	93.53	92.64	93.42	93.43
PERCENT	98.36	98.06	99.13	97.94	96.60	98.56	98.15	97.32	98.84
PERSON	98.35	98.31	98.07	98.69	98.31	100.00	98.52	98.31	99.02
TAX	91.43	82.61	84.62	88.89	90.48	73.33	90.14	86.36	78.57
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.21: Scores of BERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	64.92	70.02	57.56	75.93	81.94	69.49	69.99	75.51	62.97
COMMODITY	85.08	90.00	82.17	89.68	87.24	89.57	87.32	88.60	85.71
CURRENCY	97.31	99.18	99.54	97.95	99.18	99.32	97.63	99.18	99.43
DATE	92.92	93.55	91.69	92.92	94.82	91.92	92.92	94.18	91.80
FIN-PROD	77.25	64.62	70.27	61.72	49.41	62.90	68.62	56.00	66.38
GATHERING	71.03	86.00	79.07	83.52	82.69	87.18	76.77	84.31	82.93
INDEX	91.35	95.16	87.30	91.85	95.93	90.16	91.60	95.55	88.71
LOCATION	94.12	94.24	94.71	94.52	94.51	95.18	94.31	94.38	94.95
LOTTERY	0.00	N/A	50.00	0.00	N/A	42.86	0.00	N/A	46.15
MONETARY-VALUE	99.03	100.00	100.00	96.60	97.26	98.21	97.80	98.61	99.10
ORG	92.21	91.66	92.96	92.59	93.52	92.58	92.40	92.58	92.77
PERCENT	98.36	97.84	98.85	97.94	96.17	99.28	98.15	97.00	99.07
PERSON	98.35	98.64	97.54	98.69	98.48	99.46	98.52	98.56	98.49
TAX	94.29	79.17	45.00	94.29	90.48	64.29	94.29	84.44	52.94
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.22: Scores of BERT Base Turkish Cased per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	81.10	86.21	75.96	86.55	80.65	89.77	83.74	83.33	82.29
BUSINESS-SECTOR	63.54	72.22	57.38	78.12	85.08	70.00	70.08	78.12	63.06
COMMODITY	72.62	77.97	77.27	88.41	90.20	94.44	79.74	83.64	85.00
CURRENCY	99.14	98.37	99.55	99.14	98.77	99.32	99.14	98.57	99.44
DATE	92.55	93.78	90.70	92.79	94.55	91.16	92.67	94.17	90.93
FIN-GATHERING	25.00	100.00	N/A	50.00	50.00	N/A	33.33	66.67	N/A
FIN-ORG	92.98	93.72	92.82	90.29	93.94	88.99	91.61	93.83	90.87
FIN-PROD	75.74	67.16	71.77	60.66	51.72	71.77	67.37	58.44	71.77
INDEX	91.98	93.60	81.82	93.48	95.12	88.52	92.72	94.35	85.04
LOCATION	94.22	93.39	95.16	94.11	94.54	94.61	94.16	93.96	94.89
LOTTERY	0.00	N/A	100.00	0.00	N/A	85.71	0.00	N/A	92.31
MONETARY-VALUE	97.24	98.72	98.47	95.01	95.78	95.27	96.11	97.23	96.84
NAT-RES	89.10	94.17	92.00	85.84	85.09	87.62	87.44	89.40	89.76
NON-FIN-GATHERING	75.51	87.23	70.00	85.06	85.42	89.74	80.00	86.32	78.65
NON-FIN-ORG	85.55	86.76	86.75	88.82	88.77	88.68	87.15	87.75	87.71
PERCENT	98.28	97.22	98.42	98.11	96.60	98.99	98.19	96.91	98.71
PERSON	98.60	98.81	97.21	98.26	97.97	99.82	98.43	98.39	98.50
TAX	88.24	90.48	85.71	83.33	90.48	80.00	85.71	90.48	82.76
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.23: Scores of BERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	86.51	89.29	90.11	91.60	80.65	93.18	88.98	84.75	91.62
BUSINESS-SECTOR	62.67	70.82	56.04	75.57	83.25	69.27	68.52	76.53	61.95
COMMODITY	70.11	75.86	72.73	88.41	86.27	88.89	78.21	80.73	80.00
CURRENCY	99.02	98.97	99.09	98.81	98.97	99.09	98.92	98.97	99.09
DATE	92.51	94.85	92.15	92.27	95.37	91.92	92.39	95.11	92.04
FIN-GATHERING	22.22	40.00	N/A	50.00	50.00	N/A	30.77	44.44	N/A
FIN-ORG	93.08	94.00	92.23	92.00	95.09	87.36	92.54	94.54	89.73
FIN-PROD	75.74	69.12	70.83	61.24	55.29	68.55	67.72	61.44	69.67
INDEX	89.36	92.86	85.94	91.30	95.12	90.16	90.32	93.98	88.00
LOCATION	93.32	93.83	94.68	94.36	94.44	94.68	93.84	94.13	94.68
LOTTERY	0.00	N/A	100.00	0.00	N/A	85.71	0.00	N/A	92.31
MONETARY-VALUE	98.61	100.00	99.39	96.33	95.51	97.61	97.46	97.70	98.49
NAT-RES	90.05	94.17	83.65	86.76	85.09	82.86	88.37	89.40	83.25
NON-FIN-GATHERING	75.00	85.42	79.55	86.21	85.42	89.74	80.21	85.42	84.34
NON-FIN-ORG	85.58	87.48	86.69	88.07	88.57	89.59	86.80	88.02	88.12
PERCENT	98.20	97.62	98.85	98.20	95.96	99.28	98.20	96.78	99.07
PERSON	97.93	98.30	98.06	98.87	97.80	99.64	98.40	98.05	98.84
TAX	97.14	90.48	92.31	97.14	90.48	85.71	97.14	90.48	88.89
TIME	100.00	100.00	50.00	100.00	100.00	50.00	100.00	100.00	50.00

Table 4.24: Scores of BERT Base Turkish Cased Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	96.58	94.54	97.01	94.68	96.79	94.61
ORGANIZATION	94.17	95.01	94.60	93.61	94.39	94.31
LOCATION	96.44	96.37	96.27	95.92	96.35	96.14
DATE	-	69.53	-	76.72	-	72.95
TIME	-	100.00	-	100.00	-	100.00
MONEY	-	94.74	-	92.31	-	93.51
PERCENT	-	91.67	-	88.71	-	90.16

For FINTURK500-CHILDREN-BIO dataset, in the highest labelwise F1-scores lists for all splits, we see the entities: "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" in common. If we look at the lowest labelwise performances lists for different splits, we see NEs: "FIN-GATHERING", "FIN-PROD" and "BUSINESS -SECTOR" are common NEs in these lists. It is important to bear in mind that the NE "FIN-GATHERING" does not exist in the test set of the second split version and for this reason, only first and third split scores were taken into account when evaluating the labelwise performance for this NE.

For the FINTURK500K-CHILDREN-NON-BIO dataset, considering the labelwise versions with different split versions, the highest top 5 F1-scores lists commonly include "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" NEs. Similar to the former cases, if we compare lowest 5 labelwise F1-scores for different split versions for the same dataset, we see that "FIN-GATHERING", "FIN-PROD", "COMMODITY" and "BUSINESS-SECTOR" NEs are in common in those lists.

In addition to the datasets mentioned earlier, labelwise scores of the same model with TUR3 and EXTEND7 datasets are presented in the Table 4.24.For TUR3 dataset, the highest labelwise F1-score is observed for "PERSON" NE and the lowest labelwise F1-score is for "ORG" NE. Moreover, for the EXTEND7 dataset, which is an extended 7-label version of TUR3 dataset, the highest and lowest labelwise F1-scores are for "TIME" and "DATE" NEs, respectively.

4.5 Experiments with DistilBERT Base Multilingual Cased Model

DistilBERT Base Multilingual Cased model, which is a distilled, lighter,faster and cheaper version of BERT-Base Multilingual Cased model was presented in 2019 [3]. Model is availiable in Hugging Face library[73]. For all experiments, the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[6]. Similarly, other training arguments/parameters are set to default.

Overall performance results of experiments conducted with DistilBERT Base Multilingual Cased model using FINTURK500K BIO,FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are as shown in tables 4.25 and 4.26.

For FINTURK dataset versions (both for PARENTS and CHILDREN versions), 60% of the all datasets were used for training for split version 1. Moreover, for second and third split versions of these datasets,80% of the all datasets are used for training. For this reason, it is not interesting to observe slightly higher F1-scores for cases where more data is used for training the model.

Comparing the FINTURK's BIO and NON-BIO annotated datasets, considering the NON-BIO annotated versions of both PARENTS and CHILDREN versions of the same dataset with different split versions; performance of the model with NON-BIO datasets are slightly higher than the BIO annotated datasets. This makes sense because, learning a BIO tagging schema is a bit harder for a model with the same amount of data since in the BIO tagging schema the model also learns the beginning of the NEs.

Comparing the performance of the model using different datasets, DistilBERT Base Multilingual Cased model shows the best performance with TUR3 dataset where the F1-score using TUR3 dataset is approximately 2.5% higher than the closest, second highest, F1-score (using FINTURK500K-PARENTS-NON-BIO dataset with split version 2).This result is reasonable considering that the number of different NEs in TUR3 dataset is less than the number of different NEs in other datasets.

[6] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

In addition, considering the difference in performances using PARENTS and CHILDREN versions of FINTURK dataset, it is observed that for PARENTS versions of all dataset splits of FINTURK, overall F1-score is approximately 1% higher than the overall F1-scores of corresponding split versions of FINTURK500K-CHILDREN datasets.

Moreover, labelwise performances of DistilBERT Base Multilingual Cased model using different split versions of FINTURK500K-PARENTS-BIO dataset are shown in Table 4.27. According to the table, NEs "CURRENCY", "PERCENT", "PERSON" and "MONETARY-VALUE" are common considering the top 5 highest F1-scores lists for all splits. On the other hand, NEs "LOTTERY", "FIN-PROD" and "BUSINESS-SECTOR" are common NEs in the top 5 worst labelwise performances lists for different split versions.

Similar to the BIO annotated version, labelwise performances of the same model for FINTURK500K-PARENTS-NON-BIO dataset for different split versions are presented in Table 4.28. In the top 5 highest labelwise F1-scores lists for different split versions, NEs "PERCENT", "CURRENCY" and "PERSON" are in common. Moreover, considering the lowest 5 labelwise F1-scores lists, NEs "LOTTERY", "FIN-PROD", "GATHERING" and "BUSINESS-SECTOR" are in common.

Besides the PARENTS versions, labelwise performance results of the same model using different split versions of FINTURK-CHILDREN-BIO and FINTURK-CHILDREN-NON-BIO datasets are as shown in tables 4.29 and 4.30 respectively.For FINTURK500K-CHILDREN-BIO dataset's different split versions, the best 5 labelwise performances lists include NEs "PERCENT", "CURRENCY","PERSON" and "MONETARY-VALUE" in common. Additionally, comparing the lowest 5 labelwise F1-scores lists for different split versions, we see NEs "LOTTERY", "FIN-PROD" and "BUSINESS-SECTOR" in common.

Similar to the BIO version, for FINTURK500K-CHILDREN-NON-BIO dataset with different split versions, among the top 5 labelwise F1-score lists, "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" are in common. Moreover, taking the lowest 5 labelwise F1-score lists into account, we see "LOTTERY", "FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" take place in common.

Table 4.25: Precision, Recall and F1-Scores of DistilBERT Base Multilingual Cased Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	**FINTURK500K-PARENTS-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v3 (80-10-10)**
Precision	88.24	89.49	88.52	86.93	88.72	86.97
Recall	89.92	90.62	90.50	88.47	90.06	89.25
F1-Score	89.07	90.05	89.50	87.69	89.39	88.09

Table 4.26: Precision, Recall and F1-Scores of DistilBERT Base Multilingual Cased Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	**FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)**	**TUR3 (80-10-10)**	**EXTEND7 (80-10-10)**
Precision	88.85	90.20	89.31	87.84	89.39	87.94	93.05	89.98
Recall	90.06	90.45	90.78	88.76	90.16	89.39	92.94	89.66
F1-Score	89.45	90.32	90.04	88.30	89.77	88.66	92.99	89.82

Table 4.27: Scores of DistilBERT Base Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	58.44	67.41	52.29	72.00	79.06	68.44	64.51	72.77	59.29
COMMODITY	82.91	78.24	87.98	81.08	77.04	86.73	81.99	77.63	87.35
CURRENCY	97.84	98.56	98.64	97.32	98.16	98.20	97.58	98.36	98.42
DATE	91.40	90.84	88.89	92.12	91.83	89.34	91.75	91.33	89.11
FIN-PROD	67.82	49.35	79.44	55.92	43.68	68.55	61.30	46.34	73.59
GATHERING	65.56	66.67	67.35	64.84	69.23	84.62	65.19	67.92	75.00
INDEX	89.95	90.62	88.71	92.39	94.31	90.16	91.15	92.43	89.43
LOCATION	91.71	91.96	92.68	91.99	92.09	91.45	91.85	92.03	92.06
LOTTERY	0.00	N/A	50.00	0.00	N/A	14.29	0.00	N/A	22.22
MONETARY-VALUE	95.60	96.40	96.07	93.79	93.05	94.08	94.69	94.70	95.07
ORG	85.79	87.29	88.57	88.02	90.38	90.75	86.89	88.81	89.64
PERCENT	97.00	97.85	98.70	97.25	96.81	98.27	97.13	97.33	98.48
PERSON	94.51	95.52	96.60	95.82	93.74	96.77	95.16	94.62	96.69
TAX	78.95	90.91	85.71	83.33	95.24	80.00	81.08	93.02	82.76
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.28: Scores of DistilBERT Base Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	59.41	68.24	53.19	70.64	75.92	66.82	64.54	71.87	59.23
COMMODITY	83.68	81.25	87.00	79.36	79.59	82.46	81.46	80.41	84.67
CURRENCY	98.57	97.93	98.86	97.19	97.53	98.18	97.88	97.73	98.52
DATE	90.26	89.33	90.20	91.33	91.28	91.12	90.79	90.30	90.66
FIN-PROD	73.38	60.00	75.49	54.07	42.35	62.10	62.26	49.66	68.14
GATHERING	60.82	68.63	65.31	64.84	67.31	82.05	62.77	67.96	72.73
INDEX	90.81	91.27	87.10	91.30	93.50	88.52	91.06	92.37	87.80
LOCATION	91.72	91.73	92.43	92.08	92.13	92.35	91.90	91.93	92.39
LOTTERY	0.00	N/A	80.00	0.00	N/A	57.14	0.00	N/A	66.67
MONETARY-VALUE	97.36	97.42	98.79	95.11	94.01	97.31	96.22	95.69	98.05
ORG	86.19	88.60	90.31	89.01	89.86	90.31	87.58	89.22	90.31
PERCENT	98.45	98.48	98.71	98.11	96.81	98.99	98.28	97.64	98.85
PERSON	94.69	95.90	97.50	96.17	94.92	98.03	95.42	95.41	97.77
TAX	81.58	86.36	100.00	88.57	90.48	100.00	84.93	88.37	100.00
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.29: Scores of DistilBERT Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	81.31	55.88	87.18	73.11	61.29	77.27	76.99	58.46	81.93
BUSINESS-SECTOR	57.75	65.95	49.84	70.31	80.10	67.56	63.41	72.34	57.36
COMMODITY	61.54	82.69	65.22	81.16	84.31	83.33	70.00	83.50	73.17
CURRENCY	97.93	98.76	98.87	96.57	97.55	98.20	97.25	98.15	98.53
DATE	89.78	91.01	89.17	91.20	91.01	89.85	90.48	91.01	89.51
FIN-GATHERING	25.00	100.00	N/A	50.00	50.00	N/A	33.33	66.67	N/A
FIN-ORG	86.58	89.16	86.26	87.38	92.29	87.84	86.98	90.70	87.05
FIN-PROD	71.93	57.97	79.44	58.29	45.98	68.55	64.40	51.28	73.59
INDEX	89.36	89.15	87.50	91.30	93.50	91.80	90.32	91.27	89.60
LOCATION	90.77	91.39	92.03	91.68	92.38	92.03	91.22	91.88	92.03
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	96.08	97.11	95.76	92.71	91.56	93.49	94.37	94.25	94.61
NAT-RES	90.27	84.40	87.50	76.26	80.70	80.00	82.67	82.51	83.58
NON-FIN-GATHERING	62.11	64.00	81.58	67.82	66.67	79.49	64.84	65.31	80.52
NON-FIN-ORG	75.72	79.52	78.82	79.36	82.35	81.45	77.50	80.91	80.11
PERCENT	98.02	97.85	98.99	97.59	96.81	98.99	97.80	97.33	98.99
PERSON	94.33	95.75	97.32	95.47	95.26	97.49	94.90	95.50	97.40
TAX	78.38	82.61	85.71	80.56	90.48	80.00	79.45	86.36	82.76
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.30: Scores of DistilBERT Multilingual Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	82.83	56.52	92.11	68.91	41.94	79.55	75.23	48.15	85.37
BUSINESS-SECTOR	59.07	67.69	53.05	70.52	81.15	67.71	64.29	73.81	59.49
COMMODITY	66.67	78.57	70.83	81.16	86.27	94.44	73.20	82.24	80.95
CURRENCY	98.57	97.74	98.18	96.97	97.74	97.95	97.77	97.74	98.07
DATE	90.76	91.13	90.89	91.59	92.37	91.12	91.17	91.75	91.00
FIN-GATHERING	16.67	66.67	N/A	50.00	50.00	N/A	25.00	57.14	N/A
FIN-ORG	87.61	89.73	86.68	88.63	92.04	88.28	88.12	90.87	87.47
FIN-PROD	71.17	59.38	75.76	55.50	44.71	60.48	62.37	51.01	67.26
INDEX	90.81	91.94	91.80	91.30	92.68	91.80	91.06	92.31	91.80
LOCATION	91.79	92.24	91.91	92.47	92.64	91.60	92.13	92.44	91.76
LOTTERY	0.00	N/A	33.33	0.00	N/A	14.29	0.00	N/A	20.00
MONETARY-VALUE	98.29	97.60	99.09	93.89	91.27	97.31	96.04	94.33	98.19
NAT-RES	87.56	87.62	83.33	77.17	80.70	80.95	82.04	84.02	82.13
NON-FIN-GATHERING	67.00	77.27	82.05	77.01	70.83	82.05	71.66	73.91	82.05
NON-FIN-ORG	75.64	80.69	79.71	77.78	83.57	81.78	76.70	82.11	80.73
PERCENT	98.26	97.41	98.98	97.25	95.96	98.27	97.75	96.68	98.63
PERSON	95.08	94.90	97.65	95.91	94.42	96.95	95.49	94.66	97.30
TAX	85.71	86.96	100.00	85.71	95.24	100.00	85.71	90.91	100.00
TIME	80.00	66.67	50.00	80.00	66.67	50.00	80.00	66.67	50.00

Table 4.31: Scores of DistilBERT Base Multilingual Cased Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	93.59	88.21	93.76	89.82	93.67	89.01
ORGANIZATION	90.13	89.87	90.13	86.50	90.13	88.15
LOCATION	94.54	94.79	93.96	94.64	94.25	94.71
DATE	-	73.95	-	75.86	-	74.89
TIME	-	89.47	-	89.47	-	89.47
MONEY	-	94.44	-	87.18	-	90.67
PERCENT	-	89.09	-	84.48	-	86.73

In Table 4.31 labelwise scores of DistilBERT-Base Multilingual Cased model for TUR3 and EXTEND7 datasets are shown. Both for TUR3 and EXTEND7 datasets, the highest labelwise F1-score is associated with "LOCATION" NE. Moreover, considering TUR3 dataset, model showed the worst labelwise performance for the "ORGANIZATION" NE which also is associated with the third worst labelwise performance score considering EXTEND7 dataset. Looking at the labelwise performances using EXTEND7 dataset further, "DATE" NE is the NE which the model performed the worst.

4.6 Experiments with DistilBERT-Base Turkish Cased Model

As stated in the model's Hugging Face library webpage[73], DistilBERT Base Turkish Cased model is a community-driven model and was trained on 7GB of the original data that was used for training BERT-Base Turkish Cased model using BERT-Base-Turkish Cased model as a teacher model. For all experiments, the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[7]. Similarly, other training arguments/parameters are set to default.

Overall results of experiments conducted with DistilBERT-Base Turkish Cased model

[7] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

using FINTURK500K BIO, FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are presented in tables 4.32 and 4.33. Remembering that the 60% of the datasets are used for training for all of the first split versions whereas 80% of the datasets are used for training in the second and third split versions, it can be seen from tables that increasing the training set size affects the performance in a positive way apart from the dataset annotation type or NE variation in the dataset.

When we consider FINTURK's BIO and NON-BIO annotated versions; for same splits of BIO and NON-BIO annotated versions of the same datasets, it can be concluded that F1-scores using NON-BIO dataset versions are approximately equal or sligthly larger than the F1-scores using BIO datasets.

If we compare the model's performance using different datasets, DistilBERT-Base Turkish Cased model shows the best performance using TUR3 dataset which is the dataset with the less number of labels. F1-score using EXTEND7 dataset is approximately 3 percent lower than the former, and F1-score using FINTURK500K-PARENTS-NON-BIO dataset(with split version 2) is approximately 3.5 percent lower.

Considering the difference in performance using PARENTS and CHILDREN versions of FINTURK dataset, it is observed that for PARENTS versions of all dataset splits of FINTURK, overall F1-score is approximately 1% higher than the overall F1-scores of corresponding split versions of FINTURK500K-CHILDREN datasets.

Moreover, labelwise performances of DistilBERT-Base Turkish Cased model using different split versions of FINTURK500K-PARENTS-BIO dataset are shown in Table 4.34. As seen from the table, NEs "PERCENT", "CURRENCY" and "PERSON" are common in the top 5 labelwise F1-score lists for all splits. On the other side, if we consider the lowest 5 labelwise F1-scores lists for all splits together, we see that NEs; "LOTTERY", "GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" are in common.

Table 4.32: Precision, Recall and F1-Scores of DistilBERT Base Turkish Cased Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	**FINTURK500K-PARENTS-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v3 (80-10-10)**
Precision	88.95	89.83	89.10	88.18	89.68	88.12
Recall	90.38	91.10	91.14	89.61	90.60	90.40
F1-Score	89.66	90.46	90.11	88.89	90.14	89.24

Table 4.33: Precision, Recall and F1-Scores of DistilBERT Base Turkish Cased Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	**FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)**	**TUR3 (80-10-10)**	**EXTEND7 (80-10-10)**
Precision	89.29	90.39	89.38	88.03	89.42	88.42	93.83	90.58
Recall	90.41	91.69	91.12	89.75	90.90	90.16	94.69	91.79
F1-Score	89.85	91.03	90.24	88.88	90.15	89.28	94.26	91.18

For the NON-BIO counterpart of the former dataset, labelwise performance scores of the DistilBERT-Base Turkish Cased model are as presented in Table4.35.For three different split versions of this dataset, considering the top 5 highest F1-scores lists, we see NEs "PERCENT", "PERSON", "CURRENCY" and "MONETARY-VALUE" in common. In addition, among the 5 worst labelwise performances lists for different splits, we see NEs "LOTTERY", "FIN-PROD", "GATHERING" and "BUSINESS-SECTOR" in common.

Moreover, labelwise performance results using the FINTURK500K-CHILDREN-BIO dataset are shown in the Table 4.36. According to the table, top 5 highest F1-scores for all splits are observed in common for NEs: "PERCENT", "CURRENCY", "PERSON" and "MONETARY-VALUE". In addition, 5 worst F1-score lists for different splits include "LOTTERY", "FIN-GATHERING", "NON-FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" in common.

Furthermore, model's labelwise scores for FINTURK500K-CHILDREN-NON-BIO dataset's different split versions are as presented in Table 4.37. From the table it can be seen that, the highest top 5 labelwise F1-scores lists for different splits include "PERCENT", "CURRENCY", "MONETARY-VALUE" and "PERSON" NEs in common. Besides, the 5 lowest F1-scores lists for different splits include "LOTTERY", "FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" in common.

Moreover, labelwise scores of DistilBERT-Base Turkish Cased model with TUR3 and EXTEND7 datasets are as presented in the Table 4.38. Both for TUR3 and EXTEND7 datasets, the highest labelwise F1-score is associated with "LOCATION" NE. Moreover, considering TUR3 dataset, model showed the worst labelwise performance for the "ORGANIZATION" NE. Moreover, taking EXTEND7 dataset into consideration alone, "DATE" is the NE which the model performed the worst.

Table 4.34: Scores of DistilBERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	61.15	70.28	55.60	71.51	79.84	67.33	65.93	74.75	60.90
COMMODITY	81.25	81.12	83.77	83.05	81.12	90.52	82.14	81.12	87.02
CURRENCY	97.11	96.93	97.75	97.21	96.93	97.75	97.16	96.93	97.75
DATE	90.29	92.25	90.27	92.66	94.01	91.41	91.46	93.12	90.84
FIN-PROD	72.22	59.15	74.53	55.45	48.28	63.71	62.73	53.16	68.70
GATHERING	57.41	63.27	61.22	68.13	59.62	76.92	62.31	61.39	68.18
INDEX	85.86	88.72	90.32	89.13	95.93	91.80	87.47	92.19	91.06
LOCATION	92.47	91.45	93.20	92.68	93.03	93.20	92.58	92.23	93.20
LOTTERY	0.00	N/A	100.00	0.00	N/A	42.86	0.00	N/A	60.00
MONETARY-VALUE	97.32	96.91	96.66	93.25	93.30	94.08	95.24	95.07	95.35
ORG	85.77	86.52	89.19	89.09	89.59	90.75	87.40	88.03	89.96
PERCENT	98.27	97.20	98.41	97.42	96.17	98.13	97.84	96.68	98.27
PERSON	96.95	97.41	96.67	96.78	95.43	98.92	96.86	96.41	97.79
TAX	83.33	86.96	80.00	83.33	95.24	80.00	83.33	90.91	80.00
TIME	80.00	100.00	33.33	80.00	100.00	50.00	80.00	100.00	40.00

Table 4.35: Scores of DistilBERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	**Precision**			**Recall**			**F1-Score**		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**
BUSINESS-SECTOR	62.09	69.30	56.83	72.32	80.37	68.60	66.81	74.42	62.16
COMMODITY	84.15	82.00	86.24	84.77	83.67	89.10	84.46	82.83	87.65
CURRENCY	97.06	97.13	97.51	96.32	97.53	97.95	96.69	97.33	97.73
DATE	89.15	92.97	90.91	89.38	93.73	90.91	89.27	93.35	90.91
FIN-PROD	76.62	63.24	72.00	56.46	50.59	58.06	65.01	56.21	64.29
GATHERING	60.19	66.67	69.77	68.13	69.23	76.92	63.92	67.92	73.17
INDEX	87.96	87.69	84.38	91.30	92.68	88.52	89.60	90.12	86.40
LOCATION	92.33	91.78	93.27	92.58	93.57	93.27	92.46	92.67	93.27
LOTTERY	0.00	N/A	80.00	0.00	N/A	57.14	0.00	N/A	66.67
MONETARY-VALUE	96.53	98.71	99.09	94.57	95.51	97.01	95.54	97.08	98.04
ORG	86.81	88.26	88.25	89.42	89.87	90.62	88.09	89.06	89.42
PERCENT	98.18	97.61	98.41	97.51	95.53	97.98	97.84	96.56	98.19
PERSON	96.61	97.10	96.81	96.78	96.28	98.03	96.70	96.69	97.42
TAX	86.11	86.36	85.71	88.57	90.48	85.71	87.32	88.37	85.71
TIME	80.00	100.00	66.67	80.00	100.00	100.00	80.00	100.00	80.00

Table 4.36: Scores of DistilBERT Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	81.10	75.86	85.26	86.55	70.97	92.05	83.74	73.33	88.52
BUSINESS-SECTOR	62.11	71.50	56.35	72.12	80.10	70.00	66.74	75.56	62.44
COMMODITY	67.09	73.68	71.43	76.81	82.35	83.33	71.62	77.78	76.92
CURRENCY	97.22	97.96	98.42	97.32	97.96	98.20	97.27	97.96	98.31
DATE	90.01	93.77	89.30	92.14	94.28	90.66	91.06	94.02	89.97
FIN-GATHERING	28.57	100.00	N/A	50.00	50.00	N/A	36.36	66.67	N/A
FIN-ORG	86.19	89.50	85.33	86.59	91.38	86.70	86.39	90.43	86.01
FIN-PROD	75.00	66.15	76.47	55.45	49.43	62.90	63.76	56.58	69.03
INDEX	87.89	89.92	87.30	90.76	94.31	90.16	89.30	92.06	88.71
LOCATION	92.31	91.55	93.27	92.41	92.74	93.12	92.36	92.14	93.20
LOTTERY	0.00	N/A	100.00	0.00	N/A	57.14	0.00	N/A	72.73
MONETARY-VALUE	97.36	97.93	96.69	94.47	94.04	94.97	95.89	95.95	95.82
NAT-RES	92.82	86.09	95.00	88.58	86.84	90.48	90.65	86.46	92.68
NON-FIN-GATHERING	58.51	68.18	68.89	63.22	62.50	79.49	60.77	65.22	73.81
NON-FIN-ORG	75.88	77.91	80.18	82.36	81.11	84.79	78.99	79.48	82.42
PERCENT	98.35	97.20	98.70	97.42	96.17	98.13	97.89	96.68	98.41
PERSON	96.85	96.59	95.80	96.34	95.94	98.03	96.60	96.26	96.90
TAX	81.58	86.36	86.67	86.11	90.48	86.67	83.78	88.37	86.67
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.37: Scores of DistilBERT Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	82.54	62.50	88.17	87.39	64.52	93.18	84.90	63.49	90.61
BUSINESS-SECTOR	61.87	70.64	57.38	73.41	80.63	70.16	67.14	75.31	63.13
COMMODITY	72.97	72.88	71.43	78.26	84.31	83.33	75.52	78.18	76.92
CURRENCY	97.30	97.73	96.83	97.30	97.73	97.27	97.30	97.73	97.05
DATE	90.85	91.94	88.72	91.09	93.19	89.39	90.97	92.56	89.06
FIN-GATHERING	16.67	33.33	N/A	50.00	25.00	N/A	25.00	28.57	N/A
FIN-ORG	86.67	87.95	88.89	88.18	92.06	86.44	87.42	89.95	87.65
FIN-PROD	71.43	57.75	77.89	52.63	48.24	59.68	60.61	52.56	67.58
INDEX	89.89	92.06	87.30	91.85	94.31	90.16	90.86	93.17	88.71
LOCATION	92.88	91.68	92.81	92.66	93.07	93.27	92.77	92.37	93.04
LOTTERY	0.00	N/A	100.00	0.00	N/A	57.14	0.00	N/A	72.73
MONETARY-VALUE	97.51	98.45	98.46	95.79	95.01	95.22	96.64	96.70	96.81
NAT-RES	89.15	89.81	94.00	86.30	85.09	89.52	87.70	87.39	91.71
NON-FIN-GATHERING	66.30	72.34	66.67	70.11	70.83	76.92	68.16	71.58	71.43
NON-FIN-ORG	76.90	78.61	79.65	81.88	82.68	84.39	79.31	80.59	81.95
PERCENT	97.75	97.61	98.41	97.25	95.53	97.98	97.50	96.56	98.19
PERSON	96.84	97.27	97.00	96.08	96.62	98.39	96.46	96.94	97.69
TAX	83.33	81.82	85.71	85.71	85.71	85.71	84.51	83.72	85.71
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.38: Scores of DistilBERT Base Turkish Cased Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	94.04	90.04	95.51	92.10	94.77	91.06
ORGANIZATION	91.10	91.58	92.88	90.91	91.98	91.25
LOCATION	95.70	96.67	94.93	95.75	95.31	96.21
DATE	-	68.75	-	75.86	-	72.13
TIME	-	89.47	-	89.47	-	89.47
MONEY	-	89.47	-	87.18	-	88.31
PERCENT	-	76.06	-	87.10	-	81.20

4.7 Experiments with XLM-RoBERTa-Base Multilingual Model

In 2019[61], XLM-RoBERTa-Base Multilingual model was proposed and was described [74] as being a multilingual version of RoBERTa model[62]. Model under the Hugging Face library[74] was used during the experiments with XLM-RoBERTa-Base Multilingual model. For all experiments, the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[8]. Similarly, other training arguments/parameters are set to default.

For the XLM-RoBERTa-Base Multilingual model, overall performance results using FINTURK500K BIO,FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are given in tables 4.39 and 4.40. Comparing the splits 1&2 versions of both of the FINTURK500K-PARENTS-BIO/NON-BIO and FINTURK500K-CHILDREN-BIO/NON-BIO datasets, we may conclude that the training set size has a positive meaningful impact on the performance. However, comparing splits 1&3 together, we cannot reach such a conclusion since in some cases, the model shows a slightly worse performances using split 1 versions of both datasets. Thus,for the experiments with XLM-RoBERTa-Base Multilingual model, we cannot observe the effect of training set size on the performance.

[8] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

If we consider FINTURK dataset with BIO and NON-BIO formats and corresponding splits of these datasets, apart from split 2 versions for both PARENTS and CHILDREN datasets, we can conclude that the model performed better using the NON-BIO annotated datasets, however, we cannot generalize this observation since the situation is vice versa using the second split.

For different datasets, XLM-RoBERTa Base Multilingual model shows the best performance with using TUR3 dataset which is followed by the performance showed with using the FINTURK500K-PARENTS-BIO split 2 version. Here, it is interesting to see that the performance of the model using the EXTEND7 dataset is approximately 0.60% smaller than one of the FINTURK500K-PARENTS-BIO split versions since the EXTEND7 dataset contains only 7 labels whereas the FINTURK500K-PARENTS-BIO dataset contains 15 labels and was annotated with BIO annotation schema.

Considering the difference in performance using PARENTS and CHILDREN versions of FINTURK dataset, it is observed that for PARENTS versions of all dataset splits of FINTURK, overall F1-score is approximately 1% higher than the overall F1-scores of corresponding split versions of FINTURK500K-CHILDREN datasets.

In Table 4.41,labelwise performances of the XLM-RoBERTA Base Multilingual model using the FINTURK500K-PARENTS-BIO dataset with different split versions are shown. As it can be seen from the table, NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" are commonly included in top 5 labelwise scores lists for three splits. On the other hand, NEs "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR" and "GATHERING" are in common in the lowest 5 labelwise scores lists.

Similar to the BIO version, labelwise performances of the model on FINTURK500K-PARENTS-NON-BIO dataset is seen in Table 4.42. Considering this dataset with different three split versions, we see that NEs "PERSON", "PERCENT" and "MONETARY -VALUE" are in common in top 5 highest performances lists. For the 5 lowest performances lists, "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR" and "GATHERING" NEs exist in common.

Table 4.39: Precision, Recall and F1-Scores of XLM-RoBERTa-Base Multilingual Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	FINTURK500K-PARENTS-BIO SPLIT v1 (60-20-20)	FINTURK500K-PARENTS-BIO SPLIT v2 (80-10-10)	FINTURK500K-PARENTS-BIO SPLIT v3 (80-10-10)	FINTURK500K-CHILDREN-BIO SPLIT v1 (60-20-20)	FINTURK500K-CHILDREN-BIO SPLIT v2 (80-10-10)	FINTURK500K-CHILDREN-BIO SPLIT v3 (80-10-10)
Precision	90.16	91.68	89.69	89.07	90.73	88.58
Recall	91.33	92.33	91.61	90.54	91.66	90.51
F1-Score	90.74	92.00	90.64	89.80	91.20	89.53

Table 4.40: Precision, Recall and F1-Scores of XLM-RoBERTa-Base Multilingual Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K-PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K-PARENTS-NON-BIO SPLIT v3 (80-10-10)	FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
Precision	90.21	91.36	90.25	89.66	90.44	89.11	93.87	90.51
Recall	91.55	92.30	91.80	90.76	91.94	90.96	94.47	92.34
F1-Score	90.87	91.83	91.02	90.20	91.19	90.02	94.17	91.41

Table 4.41: Scores of XLM-RoBERTa-Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	59.88	69.37	53.82	74.64	80.63	70.44	66.45	74.58	61.02
COMMODITY	80.18	84.57	83.18	85.50	81.12	84.36	82.76	82.81	83.76
CURRENCY	98.60	98.36	98.22	98.29	98.36	99.32	98.44	98.36	98.77
DATE	92.52	93.03	91.44	92.52	94.55	91.90	92.52	93.78	91.67
FIN-PROD	67.44	59.38	78.30	54.98	43.68	66.94	60.57	50.33	72.17
GATHERING	68.57	78.85	68.09	79.12	78.85	82.05	73.47	78.85	74.42
INDEX	91.26	87.88	89.06	90.76	94.31	93.44	91.01	90.98	91.20
LOCATION	92.69	93.39	93.40	92.26	93.39	92.70	92.47	93.39	93.05
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	97.34	97.14	99.07	93.93	92.80	94.67	95.60	94.92	96.82
ORG	91.52	91.54	92.01	91.34	93.02	92.29	91.43	92.28	92.15
PERCENT	98.17	96.78	98.13	96.99	95.96	98.27	97.58	96.37	98.20
PERSON	97.66	98.47	98.40	98.00	97.97	98.92	97.83	98.22	98.66
TAX	88.57	95.45	85.71	86.11	100.00	80.00	87.32	97.67	82.76
TIME	80.00	75.00	33.33	80.00	100.00	50.00	80.00	85.71	40.00

Table 4.42: Scores of XLM-RoBERTa-Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	**Precision**			**Recall**			**F1-Score**		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**
BUSINESS-SECTOR	58.89	69.36	53.48	74.13	85.34	70.16	65.64	76.53	60.69
COMMODITY	80.05	85.05	85.31	84.77	84.18	85.31	82.34	84.62	85.31
CURRENCY	98.70	98.34	99.09	98.70	97.94	99.32	98.70	98.14	99.21
DATE	92.76	93.75	92.09	92.52	94.01	91.39	92.64	93.88	91.74
FIN-PROD	70.19	59.68	79.38	54.07	43.53	62.10	61.08	50.34	69.68
GATHERING	74.51	69.35	73.81	83.52	82.69	79.49	78.76	75.44	76.54
INDEX	89.47	91.41	93.44	92.39	95.12	93.44	90.91	93.23	93.44
LOCATION	92.92	92.75	94.84	92.74	92.35	93.19	92.83	92.55	94.01
LOTTERY	0.00	N/A	100.00	0.00	N/A	42.86	0.00	N/A	60.00
MONETARY-VALUE	98.18	97.44	99.69	95.38	95.01	96.12	96.76	96.21	97.87
ORG	91.28	91.53	91.15	91.42	92.00	93.40	91.35	91.76	92.26
PERCENT	97.75	95.94	98.70	96.82	95.53	98.41	97.28	95.74	98.56
PERSON	98.00	98.46	98.19	97.91	97.29	97.49	97.95	97.87	97.84
TAX	88.57	85.71	92.86	88.57	85.71	92.86	88.57	85.71	92.86
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.43: Scores of XLM-RoBERTa-Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	77.10	62.86	73.58	84.87	70.97	88.64	80.80	66.67	80.41
BUSINESS-SECTOR	58.45	68.10	53.53	73.20	82.72	69.11	64.99	74.70	60.33
COMMODITY	71.95	72.88	69.57	85.51	84.31	88.89	78.15	78.18	78.05
CURRENCY	98.71	97.75	99.10	98.18	97.75	99.10	98.44	97.75	99.10
DATE	91.90	93.75	90.43	92.26	94.01	90.89	92.08	93.88	90.66
FIN-GATHERING	22.22	28.57	N/A	50.00	50.00	N/A	30.77	36.36	N/A
FIN-ORG	91.96	93.30	90.97	88.55	94.17	87.84	90.22	93.74	89.38
FIN-PROD	73.62	60.00	73.87	56.87	44.83	66.13	64.17	51.32	69.79
INDEX	91.49	91.47	86.36	93.48	95.93	93.44	92.47	93.65	89.76
LOCATION	92.49	93.20	94.46	92.99	92.67	93.28	92.74	92.93	93.87
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.03	98.18	98.16	93.79	93.80	94.67	95.86	95.94	96.39
NAT-RES	90.52	89.81	91.40	87.21	85.09	80.95	88.84	87.39	85.86
NON-FIN-GATHERING	69.57	70.18	77.50	73.56	83.33	79.49	71.51	76.19	78.48
NON-FIN-ORG	81.38	86.67	81.98	85.00	85.74	86.09	83.15	86.20	83.98
PERCENT	97.07	95.31	98.13	96.82	95.11	98.27	96.95	95.21	98.20
PERSON	97.74	98.31	97.51	97.91	98.14	98.21	97.83	98.22	97.86
TAX	86.11	86.36	92.86	86.11	90.48	86.67	86.11	88.37	89.66
TIME	80.00	75.00	50.00	80.00	100.00	50.00	80.00	85.71	50.00

Table 4.44: Scores of XLM-RoBERTa-Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	71.32	75.00	79.57	77.31	67.74	84.09	74.19	71.19	81.77
BUSINESS-SECTOR	59.85	68.90	53.61	75.33	83.51	69.49	66.70	75.50	60.52
COMMODITY	67.86	68.25	70.00	82.61	84.31	77.78	74.51	75.44	73.68
CURRENCY	98.37	98.55	99.09	98.16	98.35	99.32	98.27	98.45	99.21
DATE	92.37	93.75	91.52	92.13	94.01	92.91	92.25	93.88	92.21
FIN-GATHERING	18.18	30.00	N/A	50.00	75.00	N/A	26.67	42.86	N/A
FIN-ORG	91.58	92.87	90.82	89.46	94.39	88.74	90.50	93.63	89.77
FIN-PROD	73.46	55.71	74.76	56.94	45.88	62.10	64.15	50.32	67.84
INDEX	90.48	90.15	85.94	92.93	96.75	90.16	91.69	93.33	88.00
LOCATION	93.45	91.70	95.45	93.16	92.56	94.19	93.31	92.13	94.82
LOTTERY	0.00	N/A	100.00	0.00	N/A	42.86	0.00	N/A	60.00
MONETARY-VALUE	98.03	98.19	100.00	94.43	94.76	94.93	96.19	96.45	97.40
NAT-RES	90.00	91.82	94.51	82.19	88.60	81.90	85.92	90.18	87.76
NON-FIN-GATHERING	79.76	78.85	75.68	77.01	85.42	71.79	78.36	82.00	73.68
NON-FIN-ORG	83.62	86.10	81.10	85.52	86.25	87.73	84.56	86.17	84.29
PERCENT	97.57	94.66	98.55	96.74	94.26	98.13	97.15	94.46	98.34
PERSON	97.74	98.64	97.52	98.00	98.14	98.57	97.87	98.39	98.04
TAX	100.00	100.00	92.86	100.00	100.00	92.86	100.00	100.00	92.86
TIME	80.00	33.33	50.00	80.00	33.33	50.00	80.00	33.33	50.00

Table 4.45: Scores of XLM-RoBERTa Base Multilingual Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	93.93	89.00	94.64	94.68	94.28	91.75
ORGANIZATION	91.96	91.98	93.23	90.84	92.59	91.41
LOCATION	95.28	95.38	95.20	94.18	95.24	94.78
DATE	-	68.25	-	74.14	-	71.07
TIME	-	100.00	-	100.00	-	100.00
MONEY	-	92.11	-	89.74	-	90.91
PERCENT	-	90.00	-	91.53	-	90.76

For the FINTURK500K-CHILDREN-BIO dataset with different split versions, performance results of the model are shown in Table 4.43. From the table, it can be concluded that, among the 5 most highest labelwise F1-scores lists for the three different splits of the same dataset, "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" NEs take place in common. On the other side, NEs "LOTTERY", "FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" take place in common among the 5 most lowest labelwise F1-scores.

Furhermore, for the FINTURK500K-CHILDREN-NON-BIO dataset, XLM-RoBERTa Base Multilingual model's performance results are shown in Table 4.44. Same as the BIO version of this dataset, we see NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" in common in the top 5 highest F1-scores lists for different splits. Additionally, taking lists of the 5 worst labelwise performances of the model into consideration, we see "LOTTERY", "FIN-GATHERING" and "FIN-PROD" NEs in common.

Lastly, we see the labelwise scores of the model using TUR3 and EXTEND7 datasets in Table 4.45. According to the table, for TUR3 dataset, the model performed best for "LOCATION" NE which is at the rank number two in the best labelwise performances list for EXTEND7. Moreover, "ORGANIZATION" is associated with the lowest labelwise performance of the model with TUR3. For the dataset EXTEND7, model's highest labelwise performance is for "TIME" NE. Moreover, the model can be said to be learnt the "DATE" NE worst.

4.8 Experiments with ELECTRA-Base Turkish Cased Discriminator Model

ELECTRA was proposed in 2020 [5]. ELECTRA-Base Turkish Cased Discriminator model is a community-driven ELECTRA model for Turkish. Model under the Hugging Face repository [75] was used during the experiments with ELECTRA-Base Turkish Cased Discriminator model. On the modelcard, it is stated that ELECTRA-Base Turkish Cased Discriminator model was trained on the same data as BERT-Base Turkish Cased model. Training corpus of the model has a size of 35 GB and 44,04,976,662 tokens. Moreover, the community trained a cased model on a TPU TPU v3-8 for 1M steps.

For all experiments, the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[9]. Similarly, other training arguments/parameters are set to default.

Overall performance results of ELECTRA-Base Turkish Cased Discriminator model using FINTURK500K-BIO,FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are shown in tables 4.46 and 4.47.

For splits 1&2 versions of both FINTURK500K-PARENTS-BIO/NON-BIO and FINTURK 500K-CHILDREN-BIO/NON-BIO datasets, it can be concluded that increasing the training set size increases the performance. Nonetheless this conclusion would not be generalized since for splits 1&3, we cannot reach to the same conclusion. So, for the experiments in this section, we cannot directly infer that training set size affect the performance.

For the BIO and NON-BIO versions of the FINTURK-PARENTS/CHILDREN datasets' corresponding splits, we can conclude that in general (apart from the third split of CHILDREN dataset), the model performed better using the NON-BIO annotated datasets.

[9] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

Table 4.46: Precision, Recall and F1-Scores of Electra Base Turkish Cased Discriminator Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	**FINTURK500K-PARENTS-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v3 (80-10-10)**
Precision	90.91	92.28	90.43	90.38	91.20	89.48
Recall	92.61	92.97	92.52	91.95	92.51	91.54
F1-Score	91.75	92.62	91.46	91.16	91.85	90.50

Table 4.47: Precision, Recall and F1-Scores of Electra Base Turkish Cased Discriminator Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	**FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)**	**TUR3 (80-10-10)**	**EXTEND7 (80-10-10)**
Precision	91.50	92.66	90.91	90.60	91.67	89.33	95.93	94.10
Recall	92.74	93.28	93.01	91.84	92.95	91.67	96.17	93.86
F1-Score	92.12	92.97	91.94	91.22	92.30	90.48	96.05	93.98

Comparing the performances of the model for different datasets, model performed best for TUR3 dataset which is followed by the performance for EXTEND7 and FINTURK500K-PARENTS-NON-BIO (split 2) datasets. Taking account of the different NE counts in the datasets, this result is reasonable.

Considering the difference in performance using PARENTS and CHILDREN versions of FINTURK dataset, it is observed that for PARENTS versions of all dataset splits of FINTURK, overall F1-score is approximately 0.5-1% higher than the overall F1-scores of corresponding split versions of FINTURK500K-CHILDREN datasets.

Moving on the comparison of the labelwise performances, from the table 4.48, we see that NEs: "CURRENCY", "PERSON", "PERCENT", "MONETARY-VALUE" and "LOCATION" are in common in lists of 5 most highest labelwise performances for different splits. Furthermore, NEs: "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR", "GATHERING" and "TIME" are in common in the 5 lowest labelwise performances lists for different splits.

On the other hand, for the NON-BIO version of the FINTURK500K-PARENTS dataset, as it can be clearly seen from the Table 4.49 model's top 5 labelwise performances lists for different splits contain NEs "CURRENCY", "PERSON" and "MONETARY -VALUE" in common. For the 5 lowest labelwise performances lists, "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR" and "GATHERING" are available in common.

While comparing labelwise performances of the model, considering the FINTURK500K-CHILDREN-BIO dataset, as seen in Table 4.50 we see that among the top 5 highest performances lists for 3 splits, "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" NEs are present in common. Besides, for the 5 worst labelwise performances lists, NEs "LOTTERY", "FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" take place in common.

Table 4.48: Scores of ELECTRA Base Turkish Cased Discriminator Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	61.86	72.00	55.61	76.80	84.82	71.56	68.53	77.88	62.59
COMMODITY	83.03	86.07	86.67	88.94	88.27	92.42	85.88	87.15	89.45
CURRENCY	98.93	99.18	98.88	99.04	99.18	99.32	98.98	99.18	99.10
DATE	92.04	92.70	92.17	92.40	93.46	92.17	92.22	93.08	92.17
FIN-PROD	72.00	61.43	77.36	59.72	49.43	66.13	65.28	54.78	71.30
GATHERING	70.09	83.67	66.67	82.42	78.85	82.05	75.76	81.19	73.56
INDEX	89.84	89.23	89.06	91.30	94.31	93.44	90.57	91.70	91.20
LOCATION	94.44	94.48	94.62	94.11	93.53	93.37	94.27	94.00	93.99
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.06	98.65	98.76	95.55	90.57	94.38	96.79	94.44	96.52
ORG	90.79	90.97	91.64	92.00	93.73	93.53	91.39	92.33	92.57
PERCENT	98.11	97.45	98.71	98.11	97.45	99.14	98.11	97.45	98.92
PERSON	98.52	98.46	98.75	98.26	97.46	99.46	98.39	97.96	99.11
TAX	84.21	82.61	85.71	88.89	90.48	80.00	86.49	86.36	82.76
TIME	80.00	66.67	50.00	80.00	66.67	50.00	80.00	66.67	50.00

Table 4.49: Scores of ELECTRA Base Turkish Cased Discriminator Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	61.89	70.35	56.26	74.85	85.08	71.05	67.76	77.01	62.80
COMMODITY	86.46	88.12	86.16	89.43	90.82	91.47	87.92	89.45	88.74
CURRENCY	98.71	99.17	98.20	98.92	98.76	99.09	98.81	98.97	98.64
DATE	92.72	94.31	91.48	93.45	94.82	92.17	93.08	94.57	91.82
FIN-PROD	74.53	68.66	81.19	57.42	54.12	66.13	64.86	60.53	72.89
GATHERING	67.86	78.95	65.38	83.52	86.54	87.18	74.88	82.57	74.73
INDEX	91.94	94.35	88.89	92.93	95.12	91.80	92.43	94.74	90.32
LOCATION	94.43	94.40	95.24	94.86	93.79	94.77	94.64	94.10	95.00
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	99.43	98.96	100.00	95.11	94.76	98.21	97.22	96.82	99.10
ORG	92.03	91.42	92.82	92.08	91.79	93.30	92.06	91.61	93.06
PERCENT	97.86	97.62	98.99	98.11	95.96	98.99	97.98	96.78	98.99
PERSON	98.26	98.80	98.76	98.43	97.63	100.00	98.35	98.21	99.38
TAX	97.14	100.00	72.22	97.14	100.00	92.86	97.14	100.00	81.25
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.50: Scores of ELECTRA Base Turkish Discriminator Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	80.31	87.10	89.47	85.71	87.10	96.59	82.93	87.10	92.90
BUSINESS-SECTOR	62.00	69.25	54.07	76.08	84.29	69.33	68.32	76.03	60.76
COMMODITY	63.04	75.86	66.67	84.06	86.27	88.89	72.05	80.73	76.19
CURRENCY	99.14	99.18	98.44	99.14	98.77	99.32	99.14	98.98	98.88
DATE	92.19	93.26	91.41	92.79	94.28	91.41	92.49	93.77	91.41
FIN-GATHERING	16.67	100.00	N/A	50.00	25.00	N/A	25.00	40.00	N/A
FIN-ORG	93.31	90.29	91.00	90.29	93.24	88.07	91.77	91.74	89.51
FIN-PROD	73.18	67.16	82.69	62.09	51.72	69.35	67.18	58.44	75.44
INDEX	87.11	93.70	91.80	91.85	96.75	91.80	89.42	95.20	91.80
LOCATION	94.79	93.97	95.37	93.92	93.03	93.95	94.35	93.50	94.65
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.44	98.21	99.08	93.39	95.53	95.86	95.84	96.86	97.44
NAT-RES	92.49	94.50	93.00	89.95	90.35	88.57	91.20	92.38	90.73
NON-FIN-GATHERING	68.27	89.13	68.75	81.61	85.42	84.62	74.35	87.23	75.86
NON-FIN-ORG	83.53	82.02	80.99	88.09	86.99	87.76	85.75	84.43	84.24
PERCENT	97.94	97.62	98.85	97.94	95.96	98.85	97.94	96.78	98.85
PERSON	98.27	98.13	99.11	99.04	97.63	99.46	98.66	97.88	99.28
TAX	97.14	82.61	92.86	94.44	90.48	86.67	95.77	86.36	89.66
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.51: Scores of ELECTRA Base Turkish Discriminator Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	82.03	89.29	87.23	88.24	80.65	93.18	85.02	84.75	90.11
BUSINESS-SECTOR	63.69	69.12	55.11	77.26	86.13	69.71	69.82	76.69	61.55
COMMODITY	67.47	69.84	51.72	81.16	86.27	83.33	73.68	77.19	63.83
CURRENCY	99.02	98.96	98.42	98.70	98.35	98.86	98.86	98.66	98.64
DATE	92.26	93.24	92.68	92.14	94.01	92.68	92.20	93.62	92.68
FIN-GATHERING	11.76	100.00	N/A	50.00	50.00	N/A	19.05	66.67	N/A
FIN-ORG	90.95	93.04	91.81	88.53	93.69	87.59	89.72	93.36	89.65
FIN-PROD	76.36	62.32	70.54	60.29	50.59	63.71	67.38	55.84	66.95
INDEX	89.47	93.65	91.80	92.39	95.93	91.80	90.91	94.78	91.80
LOCATION	94.06	94.61	94.51	94.21	93.72	94.44	94.13	94.16	94.47
LOTTERY	0.00	N/A	66.67	0.00	N/A	28.57	0.00	N/A	40.00
MONETARY-VALUE	99.16	99.49	99.07	95.79	97.26	95.52	97.44	98.36	97.26
NAT-RES	96.70	93.64	92.16	93.61	90.35	89.52	95.13	91.96	90.82
NON-FIN-GATHERING	78.49	83.33	75.00	83.91	83.33	84.62	81.11	83.33	79.52
NON-FIN-ORG	83.75	84.34	82.15	86.34	87.50	88.10	85.02	85.89	85.02
PERCENT	97.93	97.20	98.99	97.34	95.96	99.28	97.63	96.57	99.14
PERSON	98.18	98.30	98.24	98.52	97.97	99.82	98.35	98.14	99.02
TAX	94.44	100.00	92.86	97.14	100.00	92.86	95.77	100.00	92.86
TIME	80.00	66.67	50.00	80.00	66.67	50.00	80.00	66.67	50.00

Table 4.52: Scores of ELECTRA Base Turkish Cased Discriminator Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	97.14	95.77	97.63	96.35	97.39	96.06
ORGANIZATION	93.49	93.78	94.03	92.63	93.76	93.20
LOCATION	96.07	96.97	95.73	95.60	95.90	96.28
DATE	-	73.95	-	75.86	-	74.89
TIME	-	94.74	-	94.74	-	94.74
MONEY	-	92.31	-	92.31	-	92.31
PERCENT	-	89.06	-	91.94	-	90.48

Similar to the BIO version, performance results for FINTURK500K-CHILDREN-NON-BIO are as seen in Table 4.51. Again, we see NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" in common in lists of 5 most highest labelwise F1-scores of the model for different 3 splits. For the other side, we see "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR", "COMMODITY" and "FIN-GATHERING" NEs in common in lists of 5 lowest labelwise F1-scores for different splits.

Apart from FINTURK dataset family, the labelwise performance results using TUR3 and EXTEND7 datasets are shown in Table 4.52.Commenting on TUR3 dataset, we see that the model performed best for NE "PERSON". Second best F1-score of the model is for "LOCATION" NE which is the best learnt NE for EXTEND7. Moreover, the model's worst labelwise performances are for "ORGANIZATION" and "DATE" NEs for TUR3 and EXTEND7, respectively.

4.9 Experiments with ConvBERT-Base Turkish Cased Model

ConvBERT model was presented [6] in 2020. ConvBERT-Base Turkish Cased model is a community-driven model available under Hugging Face library[76]. As stated in the model card, model was pre-trained with 512 sequence length for 1M steps on a v3-32 TPU. Moreover, the final training corpus has a size of 35GB and 44,04,976,662

tokens.

For all experiments, the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[10]. Similarly, other training arguments/parameters are set to default.

Overall results of the experiments with ConvBERT Base Turkish Cased model using FINTURK500K-BIO,FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are presented in tables 4.53 and 4.54. Recalling that the training set size for split 1 versions of both PARENT and CHILDREN datasets are smaller than the other two split versions, we see an increase in F1-scores as the training set size increases.

Comparing the BIO and NON-BIO versions of the PARENTS and CHILDREN datasets between themselves, for the corresponding splits of the same datasets, in general, we see the performances of the model when we train using the NON-BIO datasets are slightly (approximately 0.40%) larger than training with the BIO versions (apart from the second splits of CHILDREN datasets).

The model performs best with TUR3 dataset, followed by EXTEND7 and FINTURK 500K-PARENTS-NON-BIO (split version 2 which is an expected result.

Moreover, comparing the differences in performances using PARENTS and CHILDREN versions of FINTURK dataset, it can be concluded that, in general,for PARENTS versions of different dataset splits of FINTURK, overall F1-score is approximately is slightly higher than the overall F1-scores of corresponding split versions of FINTURK500K-CHILDREN datasets.

Continuing comparisons in a labelwise manner, we see the labelwise precision, recall and F1-scores for FINTURK500K-PARENTS-BIO dataset in Table 4.55. From the table it can be said that in the top 5 highest F1-scores lists for all different splits, NEs "CURRENCY", "PERSON" and "PERCENT" are seen in common. On the other hand, NEs "FIN-PROD", "BUSINESS-SECTOR" and "GATHERING" take place in common in the 5 lowest F1-scores lists.

[10] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

Table 4.53: Precision, Recall and F1-Scores of ConvBERT Base Turkish Cased Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	FINTURK500K-PARENTS-BIO SPLIT v1(60-20-20)	FINTURK500K-PARENTS-BIO SPLIT v2(80-10-10)	FINTURK500K-PARENTS-BIO SPLIT v3(80-10-10)	FINTURK500K-CHILDREN-BIO SPLIT v1(60-20-20)	FINTURK500K-CHILDREN-BIO SPLIT v2(80-10-10)	FINTURK500K-CHILDREN-BIO SPLIT v3(80-10-10)
Precision	91.16	92.15	90.91	90.33	92.27	90.41
Recall	92.71	93.21	93.02	91.90	93.10	92.19
F1-Score	91.93	92.67	91.96	91.11	92.68	91.29

Table 4.54: Precision, Recall and F1-Scores of ConvBERT Base Turkish Cased Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	FINTURK500K-PARENTS NON-BIO SPLIT v1(60-20-20)	FINTURK500K-PARENTS NON-BIO SPLIT v2(80-10-10)	FINTURK500K-PARENTS-NON-BIO SPLIT v3(80-10-10)	FINTURK500K-CHILDREN NON-BIO SPLIT v1(60-20-20)	FINTURK500K-CHILDREN NON-BIO SPLIT v2(80-10-10)	FINTURK500K-CHILDREN NON-BIO SPLIT v3(80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
Precision	91.62	93.08	91.60	90.72	92.12	90.74	95.68	93.90
Recall	92.64	93.39	93.17	91.97	92.97	92.05	96.00	94.53
F1-Score	92.13	93.23	92.38	91.34	92.55	91.39	95.84	94.21

Table 4.55: Scores of ConvBERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	61.82	68.78	56.10	76.68	85.34	71.56	68.45	76.17	62.89
COMMODITY	87.56	85.29	90.28	89.93	88.78	92.42	88.73	87.00	91.33
CURRENCY	98.50	98.37	98.43	98.82	98.98	99.10	98.66	98.67	98.77
DATE	92.17	93.77	92.21	92.53	94.28	92.68	92.35	94.02	92.44
FIN-PROD	73.10	64.18	78.85	59.24	49.43	66.13	65.45	55.84	71.93
GATHERING	68.75	80.00	62.75	84.62	76.92	82.05	75.86	78.43	71.11
INDEX	90.96	91.27	91.80	92.93	93.50	91.80	91.94	92.37	91.80
LOCATION	94.25	93.82	95.41	94.11	93.89	94.78	94.18	93.86	95.09
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.62	99.20	96.98	96.36	92.56	94.97	97.47	95.76	95.96
ORG	91.70	92.66	92.24	91.85	93.23	94.04	91.77	92.94	93.13
PERCENT	97.93	96.79	99.14	97.68	96.17	99.42	97.81	96.48	99.28
PERSON	97.76	98.48	99.11	98.96	98.65	99.82	98.36	98.56	99.46
TAX	86.11	90.48	85.71	86.11	90.48	80.00	86.11	90.48	82.76
TIME	80.00	100.00	50.00	80.00	100.00	50.00	80.00	100.00	50.00

Table 4.56: Scores of ConvBERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	63.67	74.50	57.22	75.09	87.17	70.60	68.91	80.34	63.21
COMMODITY	86.70	90.58	90.91	89.68	88.27	90.05	88.16	89.41	90.48
CURRENCY	99.13	98.36	98.87	98.92	98.76	99.32	99.03	98.56	99.09
DATE	92.41	94.55	92.91	92.53	94.55	92.68	92.47	94.55	92.79
FIN-PROD	74.10	65.67	71.30	58.85	51.76	66.13	65.60	57.89	68.62
GATHERING	69.52	85.11	72.00	80.22	76.92	92.31	74.49	80.81	80.90
INDEX	92.35	95.12	90.32	91.85	95.12	91.80	92.10	95.12	91.06
LOCATION	94.27	94.35	95.67	94.09	94.08	95.35	94.18	94.22	95.51
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	99.03	99.74	99.69	97.15	94.76	96.72	98.08	97.19	98.18
ORG	91.49	92.37	94.02	92.28	93.21	94.02	91.89	92.79	94.02
PERCENT	97.68	96.09	98.71	97.59	94.04	99.42	97.64	95.05	99.07
PERSON	98.69	97.98	99.11	98.69	98.48	99.46	98.69	98.23	99.28
TAX	91.67	86.36	86.67	94.29	90.48	92.86	92.96	88.37	89.66
TIME	80.00	66.67	50.00	80.00	66.67	50.00	80.00	66.67	50.00

Besides, performance metrics of the model for FINTURK500K-PARENTS-NON-BIO dataset are presented in Table 4.56. For this dataset, in top 5 highest F1-scores lists for different splits, common NEs are "CURRENCY", "PERSON", "MONETARY-VALUE" and "PERCENT".On the other hand, in 5 lowest F1-scores lists, "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR", "GATHERING" and "TIME" NEs are in common.

Moving onto the FINTURK500K-CHILDREN dataset's differently annotated versions, we see the performance results of ConvBERT Base Turkish Cased model using FINTURK500K-CHILDREN-BIO dataset are as shown in Table 4.57. Here, it can be seen that NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" are the NEs which are in common in top 5 bestly learnt NEs lists. Additionally, NEs: "LOTTERY", "FIN-GATHERING", "FIN-PROD", "BUSINESS-SECTOR" and "COMMODITY" are in common in lists of 5 lowest labelwise F1-scores for various splits.

In addition, for FINTURK500K-CHILDREN-NON-BIO dataset, the 5 best scores lists for different splits include "CURRENCY", "PERSON", "MONETARY-VALUE" and "PERCENT" NEs. On the other hand, "LOTTERY", "FIN-GATHERING", "FIN-PROD", "BUSINESS-SECTOR" and "COMMODITY" are NEs that are in common in top 5 highest labelwise performances lists for different splits.

Finally, labelwise performances for TUR3 and EXTEND7 datasets are in Table 4.59. Comparing labelwise, using TUR3, model performs best for "PERSON" NE and worst for "ORGANIZATION" NE. For EXTEND7 dataset, the model performs best for "TIME" NE followed by "MONEY" NE and performed worst for "DATE".

Table 4.57: Scores of ConvBERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	83.61	88.89	94.38	85.71	77.42	95.45	84.65	82.76	94.92
BUSINESS-SECTOR	63.03	71.62	55.50	77.04	85.86	70.67	69.33	78.10	62.17
COMMODITY	64.13	78.57	63.64	85.51	86.27	77.78	73.29	82.24	70.00
CURRENCY	98.82	98.78	98.88	99.04	99.39	99.32	98.93	99.08	99.10
DATE	92.30	93.50	92.68	92.66	94.01	92.68	92.48	93.75	92.68
FIN-GATHERING	18.18	100.00	N/A	50.00	50.00	N/A	26.67	66.67	N/A
FIN-ORG	91.96	91.96	94.17	89.94	96.04	92.66	90.94	93.96	93.41
FIN-PROD	69.19	65.15	74.31	56.40	49.43	65.32	62.14	56.21	69.53
INDEX	92.97	93.55	90.32	93.48	94.31	91.80	93.22	93.93	91.06
LOCATION	94.09	94.16	95.23	93.72	93.96	94.44	93.90	94.06	94.84
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	97.67	98.98	97.86	96.09	96.53	94.67	96.87	97.74	96.24
NAT-RES	94.39	91.30	89.22	92.24	92.11	86.67	93.30	91.70	87.92
NON-FIN-GATHERING	72.55	81.63	72.92	85.06	83.33	89.74	78.31	82.47	80.46
NON-FIN-ORG	83.85	88.32	87.45	86.36	86.27	89.24	85.09	87.29	88.34
PERCENT	97.76	96.55	98.71	97.42	95.32	99.14	97.59	95.93	98.92
PERSON	97.94	98.82	99.11	99.30	98.82	99.64	98.62	98.82	99.37
TAX	88.89	100.00	76.47	88.89	100.00	86.67	88.89	100.00	81.25
TIME	80.00	75.00	50.00	80.00	100.00	50.00	80.00	85.71	50.00

Table 4.58: Scores of ConvBERT Base Turkish Cased Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	89.83	84.38	92.22	89.08	87.10	94.32	89.45	85.71	93.26
BUSINESS-SECTOR	63.30	70.32	56.47	75.33	85.60	69.93	68.79	77.21	62.49
COMMODITY	60.00	72.13	60.00	86.96	86.27	83.33	71.01	78.57	69.77
CURRENCY	98.92	98.56	99.09	98.92	98.97	99.32	98.92	98.77	99.21
DATE	92.69	93.48	92.39	93.05	93.73	91.92	92.87	93.61	92.15
FIN-GATHERING	22.22	100.00	N/A	50.00	50.00	N/A	30.77	66.67	N/A
FIN-ORG	93.20	93.38	91.51	90.50	95.56	91.72	91.83	94.46	91.62
FIN-PROD	72.50	66.67	75.47	55.50	51.76	64.52	62.87	58.28	69.57
INDEX	87.50	93.55	87.50	91.30	94.31	91.80	89.36	93.93	89.60
LOCATION	94.08	94.02	95.72	93.90	94.15	94.77	93.99	94.08	95.24
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.63	99.74	99.09	97.69	94.51	97.61	98.16	97.06	98.35
NAT-RES	93.81	97.20	97.00	89.95	91.23	92.38	91.84	94.12	94.63
NON-FIN-GATHERING	74.53	81.63	77.27	90.80	83.33	87.18	81.87	82.47	81.93
NON-FIN-ORG	84.83	87.14	87.06	87.07	87.14	87.55	85.93	87.14	87.30
PERCENT	97.42	97.82	98.84	97.16	95.53	98.41	97.29	96.66	98.63
PERSON	98.27	98.31	99.11	98.69	98.31	99.28	98.48	98.31	99.19
TAX	94.29	90.48	92.86	94.29	90.48	92.86	94.29	90.48	92.86
TIME	80.00	66.67	50.00	80.00	66.67	50.00	80.00	66.67	50.00

Table 4.59: Scores of ConvBERT Base Turkish Cased Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	96.71	94.88	97.32	95.74	97.01	95.31
ORGANIZATION	92.85	94.10	93.92	94.10	93.38	94.10
LOCATION	96.42	97.00	95.73	96.70	96.07	96.85
DATE	-	71.65	-	78.45	-	74.90
TIME	-	100.00	-	100.00	-	100.00
MONEY	-	97.44	-	97.44	-	97.44
PERCENT	-	91.67	-	88.71	-	90.16

4.10 Experiments with mDeBERTaV3 Base Multilingual Model

DeBERTa model was presented [64] in 2021 and DeBERTaV3 model[65] which brings an improvement over the former model was presented afterwards in 2021. mDeBERTaV3 which is the multilingual version of DeBERTaV3 model is available under Hugging Face library [77]. There is no structural difference between the English and multilingual versions of the model.

For all experiments, the "max_seq _length" parameter was set to 512 and model was trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[11]. Similarly, other training arguments/parameters are set to default.

Overall results of the experiments with mDeBERTaV3 Base Multilingual model using FINTURK500K BIO,FINTURK500K NON-BIO, TUR3 and EXTEND7 datasets are presented in tables 4.60 and 4.61. Since the training set size for split 1 is smaller than training set sizes for splits 2 and 3, we expect higher F1-scores for second and third splits. However, our observations does not meet with our expectations considering the splits 1 and 3 together. Thus, we cannot directly conclude that training set size has an important effect on the performance.

[11] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

Comparing the BIO and NON-BIO versions of the PARENTS and CHILDREN datasets, for the corresponding splits of the same datasets, we cannot see substantial effect of the annotation schema on the performances of the model.

The model performs best for TUR3 dataset, followed by EXTEND7 and FINTURK500K -PARENTS-NON-BIO(split version 2 which is an expected result considering the number of different NEs in these datasets.

In addition, in results of experiments with FINTURK500K-PARENTS dataset with different split versions, we see higher F1-scores compared to the results of experiments with FINTURK500K-CHILDREN dataset's corresponding split versions.

Moreover, labelwise performances of the model using FINTURK500K-PARENTS-BIO dataset are presented in Table 4.62. Considering labelwise performances, we see that among the top 5 highest labelwise F1-scores for three splits, NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" appear in common. Additionally, 5 most lowest labelwise performance metrics include NEs "LOTTERY", "FIN-PROD", "BUSINESS-SECTOR", "TIME" and "GATHERING".

Furthermore, labelwise performances for the FINTURK500K-PARENTS-NON-BIO dataset are presented in Table 4.63. From the table we see that for three splits, "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" are the NEs where the model's performance is in 5 highest performances lists in common. On the other hand, "LOTTERY", "TIME", "FIN-PROD", "BUSINESS-SECTOR" and "GATHERING" are 5 NEs which exist in common in lists of 5 lowest F1-scores.

Similar to the PARENTS case, labelwise performance metrics of model for FINTURK500K -CHILDREN-BIO dataset are as shown in Table 4.64. By looking at the table it can be deduced that, for three splits, in top 5 highest labelwise F1-scores lists; NEs "CURRENCY", "PERSON", "PERCENT" and "MONETARY-VALUE" appear in common. On the other side, among the most unsuccessfully performed NEs lists; NEs "LOTTERY", "FIN-GATHERING", "FIN-PROD" and "COMMODITY" take place in common.

Table 4.60: Precision, Recall and F1-Scores of mDeBERTaV3 Base Multilingual Model with Different Datasets and Partitions Annotated with BIO Schema

Dataset	**FINTURK500K-PARENTS-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN-BIO SPLIT v3 (80-10-10)**
Precision	90.86	91.90	91.30	90.08	90.84	89.81
Recall	92.11	93.17	92.66	91.69	92.29	91.68
F1-Score	91.48	92.53	91.97	90.88	91.56	90.73

Table 4.61: Precision, Recall and F1-Scores of mDeBERTaV3 Base Multilingual Model with Different Datasets and Partitions Annotated with NON-BIO Schema

Dataset	**FINTURK500K-PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-PARENTS-NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K-CHILDREN NON-BIO SPLIT v3 (80-10-10)**	**TUR3 (80-10-10)**	**EXTEND7 (80-10-10)**
Precision	90.75	92.17	90.73	90.32	91.23	89.84	95.33	94.02
Recall	92.43	93.22	92.35	91.75	92.63	91.73	95.75	94.07
F1-Score	91.58	92.70	91.53	91.03	91.92	90.78	95.54	94.04

Table 4.62: Scores of mDeBERTaV3 Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v1 (60-20-20)	FINTURK500K PARENTS BIO SPLIT v2 (80-10-10)	FINTURK500K PARENTS BIO SPLIT v3 (80-10-10)
BUSINESS-SECTOR	59.85	68.26	58.33	75.96	86.13	71.56	66.95	76.16	64.27
COMMODITY	84.67	85.17	82.02	88.21	90.82	88.63	86.40	87.90	85.19
CURRENCY	98.93	98.57	99.10	99.04	98.57	98.87	98.98	98.57	98.99
DATE	92.02	94.28	92.91	92.38	94.28	93.15	92.20	94.28	93.03
FIN-PROD	71.17	68.25	79.80	54.98	49.43	63.71	62.03	57.33	70.85
GATHERING	74.07	81.13	74.47	87.91	82.69	89.74	80.40	81.90	81.40
INDEX	90.48	92.19	91.80	92.93	95.93	91.80	91.69	94.02	91.80
LOCATION	93.77	94.37	95.48	92.80	93.96	94.77	93.28	94.16	95.13
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.17	96.20	98.45	93.93	94.29	94.08	96.00	95.24	96.22
ORG	92.84	92.55	93.45	92.55	93.02	93.83	92.70	92.78	93.64
PERCENT	97.42	95.26	98.57	97.34	94.04	99.14	97.38	94.65	98.85
PERSON	97.84	97.96	97.87	98.43	97.46	98.75	98.13	97.71	98.31
TAX	91.18	86.36	92.31	86.11	90.48	80.00	88.57	88.37	85.71
TIME	66.67	50.00	50.00	80.00	33.33	50.00	72.73	40.00	50.00

Table 4.63: Scores of mDeBERTaV3 Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-PARENTS Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v1 (60-20-20)**	**FINTURK500K PARENTS NON-BIO SPLIT v2 (80-10-10)**	**FINTURK500K PARENTS NON-BIO SPLIT v3 (80-10-10)**
BUSINESS-SECTOR	59.89	68.05	56.41	75.81	85.86	70.60	66.91	75.93	62.71
COMMODITY	84.60	85.29	77.97	87.71	88.78	87.20	86.13	87.00	82.33
CURRENCY	98.81	98.76	98.87	98.70	98.35	99.32	98.76	98.55	99.09
DATE	93.05	94.81	92.64	93.30	94.55	92.64	93.18	94.68	92.64
FIN-PROD	75.00	64.06	74.75	57.42	48.24	59.68	65.04	55.03	66.37
GATHERING	66.96	81.48	69.39	82.42	84.62	87.18	73.89	83.02	77.27
INDEX	90.86	92.06	93.33	91.85	94.31	91.80	91.35	93.17	92.56
LOCATION	94.06	94.18	95.77	93.66	93.43	94.10	93.86	93.80	94.93
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.19	97.95	99.69	95.92	95.51	96.12	97.04	96.72	97.87
ORG	91.15	92.23	92.46	92.64	92.70	93.61	91.89	92.46	93.03
PERCENT	97.41	97.63	98.85	97.08	96.60	98.99	97.25	97.11	98.92
PERSON	98.35	98.63	98.21	98.52	97.80	98.57	98.43	98.22	98.39
TAX	91.18	90.48	100.00	88.57	90.48	92.86	89.86	90.48	96.30
TIME	60.00	66.67	50.00	60.00	66.67	50.00	60.00	66.67	50.00

Table 4.64: Scores of mDeBERTaV3 Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	83.61	66.67	78.64	85.71	38.71	92.05	84.65	48.98	84.82
BUSINESS-SECTOR	59.51	66.94	55.11	75.96	85.86	69.56	66.74	75.23	61.49
COMMODITY	64.89	67.19	55.17	88.41	84.31	88.89	74.85	74.78	68.09
CURRENCY	98.82	98.98	98.65	98.82	98.77	99.10	98.82	98.87	98.88
DATE	92.55	93.75	92.64	93.04	94.01	92.64	92.79	93.88	92.64
FIN-GATHERING	25.00	100.00	N/A	50.00	25.00	N/A	33.33	40.00	N/A
FIN-ORG	92.80	93.71	93.02	90.97	93.93	91.74	91.88	93.82	92.38
FIN-PROD	73.17	65.67	77.14	56.87	50.57	65.32	64.00	57.14	70.74
INDEX	90.96	90.08	91.80	92.93	95.93	91.80	91.94	92.91	91.80
LOCATION	94.09	93.91	95.45	93.22	93.17	93.94	93.65	93.54	94.69
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	98.46	96.97	98.15	94.87	95.29	94.38	96.63	96.12	96.23
NAT-RES	91.59	94.44	97.89	89.50	89.47	88.57	90.53	91.89	93.00
NON-FIN-GATHERING	73.08	73.21	70.83	87.36	85.42	87.18	79.58	78.85	78.16
NON-FIN-ORG	84.96	85.03	84.72	87.82	88.06	88.50	86.37	86.51	86.57
PERCENT	97.42	95.04	98.28	97.25	93.83	98.56	97.33	94.43	98.42
PERSON	98.00	98.81	98.05	98.26	98.31	99.28	98.13	98.56	98.66
TAX	83.33	85.71	92.31	83.33	85.71	80.00	83.33	85.71	85.71
TIME	80.00	66.67	50.00	80.00	66.67	50.00	80.00	66.67	50.00

Table 4.65: Scores of mDeBERTaV3 Base Multilingual Model per Named Entity Type with Different Partitions of FINTURK500K-CHILDREN Annotated with NON-BIO Schema

Named Entities	Precision			Recall			F1-Score		
	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v3 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT v1 (60-20-20)	FINTURK500K CHILDREN NON-BIO SPLIT v2 (80-10-10)	FINTURK500K CHILDREN NON-BIO SPLIT V3 (80-10-10)
AGRIC-COMMODITY	85.22	59.38	80.81	82.35	61.29	90.91	83.76	60.32	85.56
BUSINESS-SECTOR	61.09	68.20	53.86	76.90	85.34	69.93	68.09	75.81	60.85
COMMODITY	69.32	74.14	59.26	88.41	84.31	88.89	77.71	78.90	71.11
CURRENCY	98.38	98.76	98.42	98.38	98.56	98.86	98.38	98.66	98.64
DATE	93.41	93.46	93.13	93.17	93.46	92.89	93.29	93.46	93.01
FIN-GATHERING	20.00	100.00	N/A	50.00	25.00	N/A	28.57	40.00	N/A
FIN-ORG	93.24	92.84	92.06	91.18	94.15	90.57	92.20	93.49	91.31
FIN-PROD	71.26	63.24	83.87	56.94	50.59	62.90	63.30	56.21	71.89
INDEX	90.53	93.60	93.33	93.48	95.12	91.80	91.98	94.35	92.56
LOCATION	93.98	94.48	95.69	93.43	93.94	94.01	93.70	94.21	94.84
LOTTERY	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
MONETARY-VALUE	97.38	96.95	99.69	95.79	95.01	96.72	96.58	95.97	98.18
NAT-RES	89.25	89.74	93.75	87.21	92.11	85.71	88.22	90.91	89.55
NON-FIN-GATHERING	75.76	78.43	71.74	86.21	83.33	84.62	80.65	80.81	77.65
NON-FIN-ORG	85.89	86.54	84.90	87.61	88.39	88.85	86.74	87.46	86.83
PERCENT	96.98	95.91	98.85	96.65	94.68	99.14	96.82	95.29	98.99
PERSON	98.10	98.30	97.36	98.61	97.97	99.10	98.35	98.14	98.22
TAX	91.18	90.48	92.31	88.57	90.48	85.71	89.86	90.48	88.89
TIME	80.00	66.67	33.33	80.00	66.67	50.00	80.00	66.67	40.00

Table 4.66: Scores of mDeBERTaV3 Base Multilingual Model per Named Entity Type with TUR3- EXTEND7 Datasets Annotated with NON-BIO Schema

Named Entities	Precision		Recall		F1-Score	
	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)	TUR3 (80-10-10)	EXTEND7 (80-10-10)
PERSON	96.50	94.44	96.38	95.59	96.44	95.02
ORGANIZATION	93.43	94.16	94.72	93.22	94.07	93.69
LOCATION	95.14	96.82	95.64	95.90	95.39	96.35
DATE	-	76.86	-	80.17	-	78.48
TIME	-	94.44	-	89.47	-	91.89
MONEY	-	97.30	-	92.31	-	94.74
PERCENT	-	91.53	-	93.10	-	92.31

Moreover, for FINTURK500K-CHILDREN-NON-BIO dataset, the performance metrics of the model are as shown in Table 4.65. From the table it can be concluded that, among the top 5 highest labelwise performances lists for all splits, NEs: "CURRENCY", "PERSON" and "PERCENT" appear in common. On the other side, among the 5 lowest labelwise performances lists, "LOTTERY", "FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" NEs are in common.

In addition to the labelwise performances using FINTURK datasets, we see labelwise performances for TUR3 and EXTEND7 in Table 4.66. Considering labelwise F1-scores, for TUR3, model performs the best performance for "PERSON" NE which is the NE for which the model shows the second best labelwise performance using EXTEND7. Moreover, for TUR3 and EXTEND7, model's performance is the poorest for "ORGANIZATION" and "DATE" NEs, respectively. For EXTEND7, the model performed best for "LOCATION" NE.

4.11 Experiments with 5-Fold Cross Validation Method

After the experiments with static train-dev-test partitions of FINTURK500K dataset versions, additional experiments were conducted with chosen models using K-fold cross validation method. For the experiments under this section, FINTURK500K-PARENTS-BIO dataset was used since it already had slightly higher performance scores for all models compared to FINTURK500K-CHILDREN-BIO dataset and so,

we can observe the effects of 5-fold cross validation on performance more clearly. Moreover, for these experiments, BIO annotated version was preferred since it is widely used and more popular annotation schema in the current literature.

Additionally, for all experiments, the "max_seq _length" parameter was set to 512 and models were trained with 8 batch size for 3 epochs. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[12]. Similarly, other training arguments/parameters are set to default.

The reason for these experiments is to see whether the success of the selected models are due to some coincidental issues in dataset splits or to be able to find out an improvement in the performance metrics with these more detailed experimental procedure. Thus, selecting K as 5, FINTURK500K dataset was splitted into 5 approximately approximately equal-sized partitions (sentence-wise splitting was performed) and for each experiment, for the same model, the test split was chosen as one split of total 5 splits without replacing. Here, before splitting the dataset into K partitions, the dataset, namely, sentences were shuffled first to randomize the order of sentences and then the splitting was performed.

All of the 8 models that are experimented in the scope of this thesis are also included in the experiments with K-fold cross validation. For each model, K=5 trainings and testings were performed and test performance metrics were averaged out to find mean scores for all models.

In the table 4.67 the average performance results of all models for FINTURK500K-PARENTS dataset annotated with BIO annotation schema are presented.[13] As it can be seen from the table, BERT Base Turkish Cased model outperformed all other models and it is immediately followed by ConvBERT Base Turkish Cased Model. Among the multilingual models, mDeBERTAV3 Base Multilingual Model and XLM RoBERTa Base Multilingual Cased Model showed the best performance. As expected, distilled versions of BERT multilingual and Turkish models showed the lowest performances.

Comparing the results (in Table4.68) of 5-fold cross validation experiments with ex-

[12] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html
[13] Detailed results of these experiments per 5 folds and per NE's are available in the appendix.

Table 4.67: Overall Scores of 8 Models for 5-Fold Cross Validation Experiments with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Model Name	Average Precision	Average Recall	Average F1-Score
BERT Base Multilingual Cased	92.24	93.18	92.68
BERT Base Turkish Cased	94.10	94.60	94.34
DistilBERT Base Multilingual Cased	91.30	92.28	91.78
DistilBERT Base Turkish Cased	91.44	92.58	92.00
XLM RoBERTa Base Multilingual Cased	92.88	93.60	93.26
ELECTRA Base Turkish Cased Discriminator	93.40	94.32	93.88
ConvBERT Base Turkish Cased	93.84	94.58	94.20
mDeBERTaV3 Base Multilingual	93.34	94.32	93.82

periments using static train-dev-test splits for BERT Base Multilingual Cased Model: There is an important decrease in F1-score for "TIME" when 5-fold cross validation experiments are compared to the former experiments with train-dev-test splits. Moreover, there are important increases in F1-scores for "BUSINESS-SECTOR", "COMMODITY", "FIN-PROD", "GATHERING", "LOCATION" and "TAX" NE's. F1-scores for other NE's remained approximately similar.

Comparing the results4.69 of 5-fold cross validation experiments with experiments using train-dev-test splits for BERT Base Turkish Cased Model: There are important increases in F1-Scores for "BUSINESS-SECTOR", "COMMODITY", "FIN-PROD","TAX" and "LOTTERY" NE's. Additionally, for "TIME", we see a high deviation in F1-scores obtained with 5-fold cross validation compared with results for different splits obtained for train-dev-test splits and with 5-fold cross validation we see an F1-score for this NE which is close to all F1-scores for three static splits in former experiments. Thus, we can deduce that this is since 5-fold cross validation method decreased the imbalance in distribution of NEs in test splits. Moreover, F1-Scores for other NE's did not change substantially compared to the change in the scores for the previously mentioned NE's.

For the DistilBERT Base Multilingual Cased model4.70, compared to the experiments with regular train-dev-test splits,during experiments with 5-fold cross valida-

Table 4.68: Scores of BERT Base Multilingual Cased Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	74.81	78.65	76.67
COMMODITY	89.98	92.02	90.98
CURRENCY	98.73	98.09	98.41
DATE	92.03	92.31	92.17
FIN-PROD	78.53	79.70	79.07
GATHERING	70.02	76.20	72.84
INDEX	87.29	72.87	88.85
LOCATION	95.08	95.74	95.41
LOTTERY	33.34	14.29	20
MONETARY-VALUE	96.13	95.02	95.55
ORG	91.29	92.83	92.05
PERCENT	97.96	97.72	97.84
PERSON	97.48	97.55	97.52
TAX	93.59	91.62	92.54
TIME	63.48	70.33	66.71

tion setting, we see nonignorable increases in performances of NEs: "BUSINESS-SECTOR", "COMMODITY","FIN-PROD","LOCATION" and "LOTTERY" compared to the results with 3 different splits together. Besides, with the 5-fold cross validation experiments, for "TAX" and "TIME" NEs, we see important increases considering first and third splits and decreases considering the second split since there is a strong deviation among F1-scores for different splits in former experiments.

For the DistilBERT-Base Turkish Cased model;compared to the experiments with regular train-dev-test split,during experiments with 5-fold cross validation setting where results are presented in Table 4.71, we see non-negligible increases in performances of NEs: "BUSINESS-SECTOR", "COMMODITY","FIN-PROD" and "LOCATION" compared to the results with 3 different static splits together. Besides, for "LOT-

Table 4.69: Scores of BERT Base Turkish Cased Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	79.00	80.72	79.85
COMMODITY	92.22	94.88	93.53
CURRENCY	99.38	98.61	99.00
DATE	93.50	93.57	93.53
FIN-PROD	82.08	84.19	83.10
GATHERING	76.42	81.49	78.77
INDEX	90.80	92.70	91.71
LOCATION	96.44	96.50	94.47
LOTTERY	97.14	93.14	94.92
MONETARY-VALUE	97.00	95.54	96.25
ORG	93.80	94.65	94.22
PERCENT	98.10	98.21	98.16
PERSON	98.66	99.02	98.84
TAX	95.70	95.70	95.68
TIME	71.35	77.90	73.48

TERY", "TAX" and "TIME" NEs, we see important deviations (in both increasing or decreasing side) comparing the F1-scores of experiments using different splits in train-dev-test split experiments case and with 5-fold cross validation we observed more realistic performances with more balanced distributions of NEs.

Comparing the results with the XLM-RoBERTa Base Multilingual model using 5-fold cross validation setting with experiments (results are in Table 4.72) using train-dev-test splits, we see non-negligible increases in performances of NEs: "BUSINESS-SECTOR", "COMMODITY", "FIN-PROD", "LOCATION" and "LOTTERY" considering the performances with 3 different splits together. Besides, for "TAX" and "TIME" NEs, we see important deviations (in both increasing or decreasing side) comparing the F1-scores of experiments using different splits in train-dev-test split

Table 4.70: Scores of DistilBERT Base Multilingual Cased Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	73.43	76.58	74.97
COMMODITY	88.16	90.03	89.08
CURRENCY	98.40	97.60	98.00
DATE	91.60	92.18	91.88
FIN-PROD	76.22	79.18	77.65
GATHERING	66.66	71.10	68.68
INDEX	87.29	90.52	88.85
LOCATION	94.39	95.09	94.73
LOTTERY	50.00	15.43	23.49
MONETARY-VALUE	95.14	94.04	94.56
ORG	90.03	91.55	90.78
PERCENT	98.22	97.85	98.03
PERSON	95.83	96.69	96.25
TAX	90.72	93.42	92.02
TIME	65.17	66.54	65.22

experiments case. Since the 5-fold cross validation method tackled the disadvantages arising from the imbalanced distributions in train-dev-test splits we could see more realistic results here.

Moving onto the 5-fold cross-validation experiments (results are in Table 4.73) with ELECTRA-Base Turkish Cased Discriminator model, comparing the results of these experiments with the results obtained in experiments with train-dev-test splits, it can be concluded that there are nonignorable increases in performances for NEs: "BUSINESS-SECTOR", "COMMODITY","FIN-PROD", "LOCATION" and "TAX" considering the performances with 3 different splits together. Besides, for "TIME", we see important deviations (in both increasing or decreasing side) comparing the F1-scores of experiments using different splits in train-dev-test split experiments case and

Table 4.71: Scores of DistilBERT Base Turkish Cased Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	75.06	76.94	75.98
COMMODITY	88.99	92.42	90.67
CURRENCY	98.55	97.89	98.22
DATE	91.37	92.23	91.80
FIN-PROD	73.15	76.63	74.83
GATHERING	62.49	72.15	66.94
INDEX	86.51	90.45	88.42
LOCATION	94.55	95.21	94.88
LOTTERY	50.00	18.29	26.60
MONETARY-VALUE	95.87	94.11	94.97
ORG	89.33	91.71	90.50
PERCENT	97.78	97.70	97.74
PERSON	97.30	97.53	97.41
TAX	88.93	92.54	90.62
TIME	74.92	75.90	74.59

by using 5-fold cross-validation we obtained more realistic results because of more balanced NE distribution emerging from constantly changing test-train datasets.

Continuing with the 5-fold cross-validation experiments (results are in Table 4.74) with ConvBERT-Base Turkish Cased model, comparing the results of these experiments with the results obtained in experiments with train-dev-test splits, we can see that there are substantial increases in performances for NEs: "BUSINESS-SECTOR", "COMMODITY","FIN-PROD", "GATHERING" and "TAX" considering the performances with 3 different splits together. Besides, for "TIME", we see important deviations (in both increasing or decreasing side) comparing the F1-scores of experiments using different splits in train-dev-test split experiments case and with the help of 5-fold cross validation we obtained more balanced distributions of NEs and thus

Table 4.72: Scores of XLM-RoBERTa Base Multilingual Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	76.64	79.81	78.19
COMMODITY	90.24	93.15	91.66
CURRENCY	98.76	98.38	98.57
DATE	92.24	92.54	92.39
FIN-PROD	79.19	80.42	79.72
GATHERING	75.58	81.40	78.24
INDEX	89.17	91.82	90.47
LOCATION	95.22	95.50	95.36
LOTTERY	20.00	5.71	8.89
MONETARY-VALUE	95.87	94.63	95.23
ORG	92.94	93.58	93.26
PERCENT	98.00	97.68	97.84
PERSON	98.02	98.41	98.21
TAX	93.01	95.06	94.01
TIME	68.11	73.90	70.27

obtained F1-scores that are more representative.

Lastly, comparing the results of experiments (in Table 4.75) with 5-fold-cross validation using mDeBERTaV3 Base Multilingual model to the results obtained in the classical experiments with train-dev-test splits with the same model; it can be deduced that there are important improvements in labelwise performances of the model for "BUSINESS-SECTOR", "COMMODITY", "FIN-PROD", "TAX" and "TIME". Moreover, considering the experiments with 3 different train-dev-test split cases, with mDeBERTaV3 Base Multilingual model, we see a high deviation on F1-scores for "TIME" entity. With 5-fold-cross validation setting, effect of this deviation on the performance is minimized.

Table 4.73: Scores of Electra Base Turkish Cased Discriminator Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	77.70	81.80	79.69
COMMODITY	91.76	94.35	93.03
CURRENCY	99.00	98.68	98.84
DATE	92..79	92.88	92.83
FIN-PROD	79.47	85.46	82.34
GATHERING	78.43	82.70	80.39
INDEX	88.17	91.99	90.00
LOCATION	96.10	96.12	96.11
LOTTERY	0.00	0.00	0.00
MONETARY-VALUE	96.23	95.15	95.67
ORG	92.98	94.11	93.55
PERCENT	98.02	98.06	98.04
PERSON	98.63	98.71	98.67
TAX	89.98	96.16	92.90
TIME	69.89	75.90	71.95

As a footnote for all multilingual models, during the experiments, we observed that for small number of tokens there was no prediction by the multilingual models, namely, mDeBERTaV3 Base Multilingual, BERT Base Multilingual, DistilBERT Base Multilingual and XLM-RoBERTa Base Multilingual models. To clarify, for instance for split 3, there are 50220 tokens in test dataset files for both PARENTS and CHILDREN versions (BIO/NON-BIO).However, the predicted test files include 50107 tokens for BERT Base and DistilBERT Base Multilingual models, and for mDeBERTaV3 Base Multilingual and XLM-RoBERTa Base Multilingual models, predicted test file contains 50108 and 50160 tokens, respectively. We observed that for different cases, the model stopped predicting for the current sentence when there are "girişimciliğinin" and "girişimciliğe" words. We believe that this results from

Table 4.74: Scores of ConvBERT Base Turkish Cased Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	78.46	82.12	80.24
COMMODITY	92.66	94.79	93.71
CURRENCY	99.09	98.66	98.87
DATE	93.15	93.17	93.16
FIN-PROD	80.73	85.04	82.79
GATHERING	80.44	84.40	82.32
INDEX	91.37	93.31	92.32
LOCATION	96.13	96.28	96.20
LOTTERY	0.00	0.00	0.00
MONETARY-VALUE	97.07	95.32	96.18
ORG	94.06	94.43	94.02
PERCENT	98.03	98.13	98.08
PERSON	98.82	99.06	98.94
TAX	88.71	95.35	91.82
TIME	77.72	77.71	76.99

using a multilingual model with Turkish characters since we did not observe any unpredicted tokens with monolingual Turkish pretrained models with exactly same data. Also, since the difference between tokens is very small (approximately 100 tokens are unpredicted), we can conclude that this does not affect the observed performances for these models. Besides it is important to note that we of course did not take these unpredicted tokens into account while computing the performance scores and so they did not affect anything mathematically.

Table 4.75: Scores of mDeBERTaV3 Base Multilingual Model for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Named Entities	Precision	Recall	F1-Score
BUSINESS-SECTOR	76.56	81.43	78.91
COMMODITY	91.22	93.90	92.53
CURRENCY	99.03	98.49	98.76
DATE	92.65	92.86	92.75
FIN-PROD	78.92	81.84	80.32
GATHERING	76.44	79.78	78.05
INDEX	90.16	92.42	91.27
LOCATION	96.13	96.24	96.18
LOTTERY	0.00	0.00	0.00
MONETARY-VALUE	96.37	95.38	95.86
ORG	93.39	94.41	93.90
PERCENT	98.12	98.14	98.13
PERSON	98.46	98.77	98.61
TAX	93.64	95.87	94.62
TIME	71.80	76.30	73.46

4.12 Experiments with FINTURK500K-PARENTS-BIO Dataset Using Different Language-Specific BERT Models

As we presented among the results above, BERTurk model which was trained with a huge amount of Turkish data shows a very high success compared to multilingual BERT Base model. Considering the Ural-Altaic language group,since languages under this group are not close relatives but are much more closer to each other considering languages in other language families, it would not be a big surprise if a language specific model for a language in this group, provides a good performance on Turkish texts. Actually, this kind of a study was conducted before [8] where authors compared transferability of different Uralic languages and they even experimented with

Table 4.76: Overall Scores of Language Specific BERT Models using FINTURK500K-PARENTS Dataset Annotated with BIO Schema

Model Name	Trained Language	Precision	Recall	F1-Score
BERT Base Finnish Cased V1 Model	Finnish	82.73	81.92	82.33
EstBERT Estonian Cased Model	Estonian	80.44	80.85	80.64
huBERT Hungarian Cased Model	Hungarian	85.06	87.57	86.30
UzBERT Base Uncased Model	Uzbek	82.74	84.64	83.68
BERT Base Cased Model	English	86.99	89.18	88.07
KR-BERT Model	Korean	83.25	86.95	85.06
GREEK-BERT Uncased V1 Model	Greek	81.19	83.21	82.19

languages under this language group with Cyrillic alphabet.

We used language specific BERT[14] models for Finnish[22][78], Hungarian[24][79], Estonian[23][80], Uzbek[49][81] and EngBERT (English BERT)[2][82] from Huggingface platform and trained (fine-tuned) using our FINTURK500K-PARENTS-BIO train dataset with these models with 3.0 epochs, setting maximum sequence length parameter as 512 and batch size as 8, and we obtained very high F1-scores. Moreover, the initial learning rate parameter was set to default, namely 5e-05 which is the initial learning rate for Adam optimizer defined in Trainer class[15]. Similarly, other training arguments/parameters are set to default.

Afterwards, to see the performance of monolingual BERT models trained with languages in different alphabets/writing styles, we experimented with KR-BERT [25][83] and GREEK-BERT Uncased V1 [26][84] models based on BERT architecture and as seen on the Table 4.76 we again obtained good results using our Turkish dataset.

High results with Finnish, Hungarian and Estonian is not as surprising as successes with Uzbek BERT, Greek BERT and Korean BERT models since these models were trained with data in Cyrillic, Greek and Hangul writing systems, respectively.

[14] Actually BERT-based models will be a more precise since in some of the related articles authors mention that they trained a language specific model for their language using the same/similar architecture as BERT-Base or in some of the studies the base model is not explicitly stated but we deduced that BERT-Base model, namely the original first BERT-Base (English) model or its variants were used as a base architecture.

[15] https://huggingface.co/transformers/v3.5.1/main_classes/trainer.html

The success of the Uzbek (with Cyrillic alphabet)-specific BERT model on Turkish texts is interesting since Turkish uses Latin alphabet as the writing system. Since Uzbek which is also an Altaic and Turkic language such as Turkish, with also an agglutinative language, constructing many different words from a single root via adding affixes is usual. Despite the fact that two writing systems are very different from each other, the since the morphology is similar, it may be the case that UzBERT model performed well on Turkish even if it was trained with Cyrillic alphabet. Another similar case is with Korean where we observed a 85.06% as F1-score despite the difference in Hangul and Latin alphabets. In literature survey chapter, we mentioned that it is known that there are some linguists who classify Korean in Altay language family as well. Thus, it may be thought that one reason for the success may be the kinship between Turkish and Korean.

However, another interesting thing is the success with the English BERT. English is under Indo-European language family which is not close to Turkish at all. There are some fundamental differences between Turkish and English which is a morphologically poorer language:

- The order of items in a sentence: In English the order of items is "subject-verb-object" whereas in Turkish the order is generally "subject-object-verb" which can be changed due to inverted sentences
- Number of tenses: As mentioned in [85], in Turkish there are 5 tenses fundamentally whereas this number increases to 12 when it comes to English. Some tenses in English does not have Turkish equivalents such as the present perfect tense. Also again as remarked in [85], English has regular and irregular verbs whereas the Turkish language uses fixed suffixes with all tenses associated with the past, present, and future.
- Usage of the subject: In an English sentence the subject has to be used whereas in a Turkish sentence, the meaning may not change even if the subject may not be used. For example, in the sentence: "Ben okula gidiyorum." (meaning: "I am going to school.") can be said like "Okula gidiyorum." (without using the "Ben" (meaning:"I") word without any loss in meaning. However, the sentence "I am going to school" does not preserve the meaning if we erase the "I am"

part.

- Plural forms of some words: In English, there are some uncountable nouns such as money, salt and water. However, in Turkish there is not such a concept, plural forms of words can be constructed by using suffixes such as "saç-saçlar" (hair (singular)-hair (plural)).

Since English-specific BERT model provided high scores with Turkish dataset, we can conclude that the success in cross-lingual transfer does not benefit from language relatedness. Our conclusion is aligned with the conclusion that authors of [8] reached, they received high results with EngBERT on Finnish, Hungarian and Estonian datasets. To be more specific, despite the fact that Finnish, Hungarian and Estonian are under same language group, EngBERT performed better than EstBERT and huBERT using Finnish data and also performed better than EstBERT and FinBERT models using Hungarian data. Moreover, with Estonian data; there is no substantial gap between the performances of EngBERT, FinBERT and huBERT. However, they did not trained (trained to fine-tune) a dataset with latin alphabet using a pre-trained model for a Cyrillic alphabet. They compared the performances of models trained for Cyrillic languages using Cyrillic datasets. For instance, they received good performances with RuBERT (Russian BERT) on datasets in Erzya, Moksha, Komi Permyak and Komi Zyrian (languages written in Cyrillic).

In classical rule-based or traditional machine learning techniques, cross-lingual transfer can not be very successful since the features that models utilize are handcrafted or some domain-specific resources such as gazetteers and lexicons are used in the modeling process. In this regard, as stated in [36], *"The characteristics of the source language to extract information from also have a significant impact on the extraction techniques being used. A certain feature of one language, which can help the extraction process, may not be available for another"*. However, this deduction is not applicable for deep-learning based methods as we see relatively good results in cross-lingual transfer experiments using languages that are classified in different language groups or even using different writing systems-alphabets.

At the end of the this section, it is important to note that using many various Hugging Face models in all experiments, to fine-tune library models for token classification,

we benefited from some code from Github[16][17] [18] copyrighted by Google AI Language Team Authors, The HuggingFace Inc. team and NVIDIA.

[16] https://github.com/huggingface/transformers/blob/main/examples/pytorch/token-classification/run_ner.py
[17] https://github.com/huggingface/transformers/blob/main/examples/legacy/token-classification/utils_ner.py
[18] https://github.com/huggingface/transformers/blob/main/examples/legacy/token-classification/scripts/preprocess.py

CHAPTER 5

CONCLUSIONS

To sum up, in this study, we experimented most popular multilingual or monolingual deep-learning based models with mainly 4 datasets where 2 of these are newly presented and annotated using two different schemas. Additionally, since we had 2 differently (BIO/NON-BIO) annotated versions of same datasets, we trained the models for both formats and compared the successes using the exact same tool after converting the test predictions of NON-BIO datasets into BIO format. This way, we had an opportunity to observe the effect of annotation schema on the performance. We also took a look at the language-specific BERT models for languages in the same language group with Turkish and compared the results of these monolingual models on one of our datasets.

As it can be seen from the results of the experiments provided in 5, Microsoft's multilingual mDeBERTaV3 model (mDeBERTaV3) showed a noticeably good performance considering it is a multilingual model.

Furthermore, as expected, the best performance is obtained with models pre-trained for Turkish. Among the monolingual models, considering all datasets generally, BERT Base Turkish Cased model performed best which is immediately followed by ConvBERT Base Turkish Cased model. However, considering the slight difference obtained using best monolingual and multilingual models (especially mDeBERTaV3), it can be concluded that the need for training the model with massive amount of Turkish corpora is not as important as before considering the most currently developed multilingual models. In other words, in some cases where the cost and time is more important, apart from training a language specific model, one can continue with a successful multilingual model.

In addition, it is also worth to re-mention that, for TUR3 and EXTEND7 datasets, F1-scores with BERT Base Turkish Cased model are 96.07 and 93.58, respectively. Comparing these with scores obtained with same datasets in former studies in the literature, to the best of our knowledge, we observed the highest F1-score for EXTEND7 and second highest F1-score for TUR3 dataset. The highest F1-score (96.48) for TUR3 was obtained by Safaya et al. [86] in March 2022 by adding a CRF layer on the top of BERTurk model by[87].

Moreover, we experimented with 5-fold cross validation method using FINTURK500K-PARENTS-BIO dataset and observed that the performance scores were increased for all models, in general. This conclusion meets with our expectations since with k-fold cross validation, the test and training datasets change periodically and this makes the model see much more different examples during training and so, learn better.

The dataset we presented includes many different tags related to financial news. However, after observing the results of the experiments we conclude that using both FINTURK500K-PARENTS and FINTURK500K-CHILDREN datasets success of models on "LOTTERY" NE is very poor due to the scarcity of this named entity in the text compared to the other named entities. Indeed, "LOTTERY" is not a concept that is directly related to financial sector. However, since in financial news texts was our corpus, before annotating the dataset, we came across news related to "LOTTERY" and we thought that including this named entity will increase the diversity of named entities since the relevant news are include monetary values, currencies and so on. Besides the "LOTTERY" NE, "FIN-PROD", "GATHERING", "TIME" and "BUSINESS-SECTOR" are four other NEs in the FINTURK500K-PARENTS dataset, which we saw most lowest performances in approximately all experiments with various models. For FINTURK500K-CHILDREN dataset, we see relatively lower labelwise performances for "FIN-GATHERING", "FIN-PROD" and "BUSINESS-SECTOR" besides "LOTTERY".

Considering the NEs that are not well-predicted by different models, it can be concluded that since the counts for these specific entities are relatively smaller than the others, models could not successfully learn these NEs. Another interpretation related to labelwise results may be that; the NE "TAX" is not well represented in the dataset,

besides we see a pretty successful prediction related to "TAX", among other labelwise performances. The reason for this success may be the fact that the variation in "TAX" NEs is very limited, i.e. there are a few certain tax types that we frequently face with when it comes to financial news texts. Thus, besides the count of an NE, the variation related to that NE in a dataset is very determinant on the model's performance since the model may memorize that NE after some point.

Furthermore, about the "MONETARY-VALUE", "CURRENCY" NEs from FINTURK, and "MONEY" from EXTEND7 datasets: With all models, we see very high performances for "CURRENCY" and "MONETARY-VALUE" NEs, one reason for this is that there is a pattern since in general, a monetary value is preceded by a currency unit. This pattern eases the learning process for all models. Moreover, in EXTEND7, "MONEY" NE contains both the monetary-value and currency whereas in FINTURK datasets, these are two different NEs. Taking all experiments with different models into account, we can conlude that the entitywise performance for "MONETARY-VALUE" is higher than the entitywise performance for "MONEY". Therefore, splitting monetary-amount from currency as two distinct NEs eases the learning process for monetary expression.

Another conclusion that we derive from the studies in the scope of this thesis work is related to our observations of similar performances when we used language-specific BERT models with our FINTURK500K-PARENTS dataset. This emphasizes that the performance of the BERT model strongly depends on the model's internal dynamics rather than the language it is pre-trained. In addition, when the relativeness of languages is taken into account, we can conlude that the closeness of languages has approximately no substantial effect on the performance when applying a model pre-trained with datasets in a language, on a dataset in another language.

5.1 Future Work

Future work based on this thesis comprises nested NER studies with the FINTURK dataset's related version. Deep learning-based techniques may provide considerable success in nested NER field as well. Combining some of the NEs in FINTURK500K-

PARENTS and FINTURK500K-CHILDREN datasets in a nested fashion considering the hierarchical relations between some of the NEs such as "COMMODITY" and "AGRIC-COMMODITY" (agricultural commodity)/ "NAT-RES" (natural resource) or adding some new NEs such as "TITLE" which may be a part of a higher NE "PERSON", a newer dataset can be created containing both nested NEs and flat NEs together.

Another future plan related to the further study of these subjects, we aim to experiment on further deep-learning based models using both financial news texts and annotating Web 2.0 data (such as tweets in finance domain) which is more informal and ungrammatical, as well.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019.

[4] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *CoRR*, vol. abs/1901.07291, 2019.

[5] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *International Conference on Learning Representations*, 2020.

[6] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, "Convbert: Improving BERT with span-based dynamic convolution," *CoRR*, vol. abs/2008.02496, 2020.

[7] P. P. He, X. Liu, J. Gao, and W. Chen, "Microsoft deberta surpasses human performance on the superglue benchmark," Jan 2021.

[8] J. Ács, D. Lévai, and A. Kornai, "Evaluating transferability of BERT models on uralic languages," *CoRR*, vol. abs/2109.06327, 2021.

[9] B. Indig and I. Endrédy, *Gut, Besser, Chunker – Selecting the Best Models for Text Chunking with Voting*, pp. 409–423. 01 2018.

[10] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. D. Spyropoulos, and P. Stamatopoulos, "Rule-based named entity recognition for greek finan-

cial texts," in *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pp. 75–78, Citeseer, 2000.

[11] G. Tür, D. Hakkani-Tür, and K. Oflazer, "A statistical information extraction system for turkish," *Natural Language Engineering*, vol. 9, no. 2, pp. 181–210, 2003.

[12] S. Ozkaya and B. Diri, "Named entity recognition by conditional random fields from turkish informal texts," *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, pp. 662–665, 2011.

[13] N. Chinchor and P. Robinson, "Appendix E: MUC-7 named entity task definition (version 3.5)," in *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998.

[14] E. T. K. Sang, "Introduction to the conll-2002 shared task: Language-independent named entity recognition," *ArXiv*, vol. cs.CL/0209010, 2002.

[15] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[16] D. Küçük and A. Yazici, "A hybrid named entity recognizer for turkish with applications to different text genres," vol. 62, pp. 113–116, 01 2010.

[17] R. Yeniterzi, "Exploiting morphology in Turkish named entity recognition system," in *Proceedings of the ACL 2011 Student Session*, (Portland, OR, USA), pp. 105–110, Association for Computational Linguistics, June 2011.

[18] C. Tosun, "Dil zenginliği, yozlaşma ve türkçe," *Journal of Language and Linguistic Studies*, vol. 1, no. 2, pp. 136–154, 2005.

[19] "Türk dil kurumu sözlükleri." `https://sozluk.gov.tr/`. Accessed on: 2010-08-11, from the menu: "Atasözleri ve Deyimler Sözlüğü".

[20] D. Aksan, "Türk dili zengin bir dil midir?," *Atatürk ve Türk Dili ve edebiyatı, Türk eğitimi ve Türk kültürü konusunda seçme yazılar*, p. 84, 1972.

[21] "Turkish BERT Model(with Vocabulary Size 128k) on HuggingFace Library." https://huggingface.co/dbmdz/bert-base-turkish-128k-cased. Accessed on 12.07.2022.

[22] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo, "Multilingual is not enough: Bert for finnish," *arXiv preprint arXiv:1912.07076*, 2019.

[23] H. Tanvir, C. Kittask, S. Eiche, and K. Sirts, "Estbert: A pretrained language-specific bert for estonian," 2020.

[24] N. D. Márk, "Introducing hubert," https://hlt.bme.hu/media/pdf/huBERT.pdf, Accessed on 13.07.2022.

[25] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "Kr-bert: A small-scale korean-specific language model," 2020.

[26] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos, "GREEK-BERT: The greeks visiting sesame street," in *11th Hellenic Conference on Artificial Intelligence*, ACM, sep 2020.

[27] "German BERT-Base Model." https://huggingface.co/bert-base-german-cased. Accessed on 12.07.2022.

[28] W. Antoun, F. Baly, and H. M. Hajj, "Arabert: Transformer-based model for arabic language understanding," *CoRR*, vol. abs/2003.00104, 2020.

[29] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[30] T. N. Gencan, "Dilbilgisi, türk dil kurumu yayınları, 4," *Baskı, Ankara*, vol. 352, 1979.

[31] P. Yalçın, "Turkish language and its place among world languages," *Dil Dergisi*, vol. 47, pp. 30–35, 1996.

[32] A. B. Ercilasun, "The place of turkish among the world languages," *Dil Araştırmaları*, vol. 12, pp. 9–16, 2013.

[33] S. Cucerzan and D. Yarowsky, "Language independent named entity recognition combining morphological and contextual evidence," in *1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, 1999.

[34] O. Bayraktar and T. Taskaya Temizel, "Person name extraction from turkish financial news text using local grammar based approach," pp. 1 – 4, 11 2008.

[35] D. Küçük and A. Yazici, "Named entity recognition experiments on turkish texts," pp. 524–535, 10 2009.

[36] S. Tatar and I. Cicekli, "Automatic rule learning exploiting morphological features for named entity recognition in turkish," *J. Information Science*, vol. 37, pp. 137–151, 04 2011.

[37] G. A. Şeker and G. Eryiğit, "Initial explorations on using CRFs for Turkish named entity recognition," in *Proceedings of COLING 2012*, (Mumbai, India), pp. 2459–2474, The COLING 2012 Organizing Committee, Dec. 2012.

[38] B. Eken and A. Tantuğ, "Recognizing named entities in turkish tweets," vol. 5, pp. 155–162, 01 2015.

[39] G. A. Seker and G. Eryigit, "Extending a crf-based named entity recognition model for turkish well formed text and user generated content," *Semantic Web*, vol. 8, pp. 625–642, 2017.

[40] G. Celikkaya, D. Torunoğlu, G. Eryiğit, and D. Torunoğlu-Selamet, "Named entity recognition on real data: A preliminary investigation for turkish," 10 2013.

[41] H. Demir and A. Özgür, "Improving named entity recognition for morphologically rich languages using word embeddings," *2014 13th International Conference on Machine Learning and Applications*, pp. 117–122, 2014.

[42] M. Ševčíková, Z. Žabokrtskỳ, and O. Krza, "Named entities in czech: annotating data and developing ne tagger," in *International Conference on Text, Speech and Dialogue*, pp. 188–195, Springer, 2007.

[43] O. Kuru, O. A. Can, and D. Yuret, "CharNER: Character-level named entity recognition," in *Proceedings of COLING 2016, the 26th International Confer-*

ence on Computational Linguistics: Technical Papers, (Osaka, Japan), pp. 911–921, The COLING 2016 Organizing Committee, Dec. 2016.

[44] O. Güngör, T. Gungor, and S. Uskudarli, "The effect of morphology in named entity recognition with sequence tagging," *Natural Language Engineering*, vol. 25, pp. 147–169, 01 2019.

[45] Y. Kilic, D. Dinc, and P. Karagoz, "Named entity recognition on morphologically rich language: Exploring the performance of bert with varying training levels," pp. 4613–4619, 12 2020.

[46] E. K. Akkaya and B. Can, "Transfer learning for turkish named entity recognition on noisy text," *Natural Language Engineering*, vol. 27, no. 1, pp. 35–64, 2021.

[47] G. Aras, D. Makaroglu, S. Demir, and A. Cakir, "An evaluation of recent neural sequence tagging models in turkish named entity recognition," *CoRR*, vol. abs/2005.07692, 2020.

[48] X. Cheng, W. Wang, F. Bao, and G. Gao, "Mtner: A corpus for mongolian tourism named entity recognition," in *China Conference on Machine Translation*, pp. 11–23, Springer, 2020.

[49] B. Mansurov and A. Mansurov, "Uzbert: pretraining a BERT model for uzbek," *CoRR*, vol. abs/2108.09814, 2021.

[50] I. Demiros, S. Boutsis, V. Giouli, M. Liakata, H. Papageorgiou, and S. Piperidis, "Named entity recognition in greek texts," in *LREC*, 2000.

[51] S. Wang, R. Xu, B. Liu, L. Gui, and Y. Zhou, "Financial named entity recognition based on conditional random fields and information entropy," in *2014 International Conference on Machine Learning and Cybernetics*, vol. 2, pp. 838–843, IEEE, 2014.

[52] E. Emekligil, S. Arslan, and O. Agin, "A bank information extraction system based on named entity recognition with crfs from noisy customer order texts in turkish," in *International Conference on Knowledge Engineering and the Semantic Web*, pp. 93–102, Springer, 2016.

[53] S. Francis, J. Van Landeghem, and M.-F. Moens, "Transfer learning for named entity recognition in financial and biomedical documents," *Information*, vol. 10, no. 8, 2019.

[54] D. Sagheer and F. Sukkar, "Text template mining using named entity recognition," *International Journal of Computer Applications*, vol. 182, pp. 34–40, 03 2019.

[55] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "Finbert: A pre-trained financial language representation model for financial text mining," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (C. Bessiere, ed.), pp. 4513–4519, International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.

[56] A. Jabbari, O. Sauvage, H. Zeine, and H. Chergui, "A French corpus and annotation schema for named entity recognition and relation extraction of financial news," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 2293–2299, European Language Resources Association, May 2020.

[57] L. Zhao, L. Li, and X. Zheng, "A bert based sentiment analysis and key entity detection approach for online financial texts," 2020.

[58] O. KABASAKAL and A. MUTLU, "Named entity recognition in turkish bank documents," *Kocaeli Journal of Science and Engineering*, vol. 4, no. 2, pp. 86–92.

[59] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus," *CoRR*, vol. abs/2009.05451, 2020.

[60] R. Hanslo, "Deep learning transformer architecture for named entity recognition on low resourced languages: State of the art results," 11 2021.

[61] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019.

[62] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[63] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," 2019.

[64] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021.

[65] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," 2021.

[66] "Python natural language toolkit web page." `https://www.nltk.org/`. Accessed on: 2010-12-29.

[67] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960. `https://w3.ric.edu/faculty/organic/coge/cohen1960.pdf`, Accessed on 13.07.2022.

[68] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: Measures of agreement," *Perspectives in clinical research*, vol. 8, no. 4, p. 187, 2017.

[69] R. Jiang, R. E. Banchs, and H. Li, "Evaluating and combining name entity recognition systems," in *Proceedings of the Sixth Named Entity Workshop*, pp. 21–27, 2016.

[70] N. Alshammari and S. Alanazi, "The impact of using different annotation schemes on named entity recognition," *Egyptian Informatics Journal*, vol. 22, 11 2020.

[71] "BERT Base Multilingual Cased Model." `https://huggingface.co/bert-base-multilingual-cased`. Accessed on 12.07.2022.

[72] S. Schweter, "Berturk - bert models for turkish," Apr. 2020.

[73] "DistilBERT Base Multilingual Cased Model." https://huggingface.co/distilbert-base-multilingual-cased. Accessed on 12.07.2022.

[74] "XLM-RoBERTa Base Multilingual Cased Model." https://huggingface.co/xlm-roberta-base. Accessed on 12.07.2022.

[75] "ELECTRA Base Turkish Cased Discriminator Model." https://huggingface.co/dbmdz/electra-base-turkish-cased-discriminator. Accessed on 12.07.2022.

[76] "ConvBERT Base Turkish Cased Model." https://huggingface.co/dbmdz/convbert-base-turkish-cased. Accessed on 12.07.2022.

[77] "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing." https://huggingface.co/microsoft/mdeberta-v3-base. Accessed on 12.07.2022.

[78] "BERT Base Finnish CasedV1 Model." https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1. Accessed on 12.07.2022.

[79] "huBERT Base Model(Cased)." https://huggingface.co/SZTAKI-HLT/hubert-base-cc. Accessed on 12.07.2022.

[80] "EstBERT." https://huggingface.co/tartuNLP/EstBERT. Accessed on 12.07.2022.

[81] "UzBERT Base Model(Uncased)." https://huggingface.co/coppercitylabs/uzbert-base-uncased. Accessed on 12.07.2022.

[82] "BERT Base Model(Cased)." https://huggingface.co/bert-base-cased. Accessed on 12.07.2022.

[83] "KoRean Based BERT pre-trained (KR-BERT)." https://huggingface.co/snunlp/KR-BERT-char16424. Accessed on 12.07.2022.

[84] "GreekBERT." https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1. Accessed on 12.07.2022.

[85] N. Jomaa, "The perspectives of turkish efl learners on the differences and similarities between turkish (l1) and english (l2) languages," *Language Teaching and Educational Research*, vol. 4, no. 2, pp. 148 – 160, 2021.

[86] A. Safaya, E. Kurtuluş, A. Göktoğan, and D. Yuret, "Mukayese: Turkish nlp strikes back," 03 2022.

[87] S. Schweter, "BERTurk - BERT models for Turkish." `https://zenodo.org/record/3770924#.YvkwNRxBw5k`. Accessed on 26.07.2022.

APPENDIX A

DETAILED RESULTS OF 5-FOLD CROSS VALIDATION EXPERIMENTS FOR DIFFERENT MODELS PER NE AND FOR 5 DIFFERENT FOLDS

During the 5-fold cross-validation experiments, FINTURK500K-PARENTS-BIO was splitted into sentences and shuffled randomly. Afterward, the dataset was splitted into 5 different partitions. Experiments with all 8 models were performed with 5 folds, each time setting one of the folds as test set and combining other folds as training set. Overall results were presented in "Experiments and Results" chapter and detailed results for each fold and each model are as shown in tables below.

Table A.1: Scores of BERT Models for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

		BERT Base Multilingual Cased			**BERT Base Turkish Cased**		
Fold	**Named Entity**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
1	BUSINESS-SECTOR	73.22	76.92	75.02	79.21	80.97	80.08
2		75.78	78.67	77.20	79.90	80.53	80.21
3		75.32	81.04	78.08	79.41	83.13	81.23
4		75.44	81.08	78.16	79.56	82.12	80.82
5		74.28	75.55	74.91	76.92	76.85	76.89
Average		**74.81**	**78.65**	**76.67**	**79.00**	**80.72**	**79.85**
1	COMMODITY	90.97	91.86	91.41	92.50	94.46	93.47
2		89.52	89.68	89.60	91.72	94.66	93.17
3		90.46	93.06	91.74	94.11	94.59	94.35
4		89.05	92.37	90.68	92.50	96.10	94.27
5		89.89	93.12	91.48	90.27	94.59	92.38
Average		**89.98**	**92.02**	**90.98**	**92.22**	**94.88**	**93.53**

Table A.1: Continued

1	CURRENCY	98.80	97.74	98.27	99.40	98.57	98.99
2		99.28	97.75	98.51	99.52	98.22	98.87
3		97.88	98.76	98.32	99.43	98.99	99.21
4		98.75	98.19	98.47	99.09	98.64	98.87
5		98.95	98.03	98.49	99.48	98.65	99.06
Average		**98.73**	**98.09**	**98.41**	**99.38**	**98.61**	**99.00**

1	DATE	92.15	92.62	92.39	91.90	92.37	92.13
2		92.98	92.98	92.98	94.47	94.08	94.27
3		92.45	93.28	92.86	95.50	95.75	95.62
4		91.09	90.71	90.90	91.82	91.82	91.82
5		91.47	91.96	91.71	93.83	93.83	93.83
Average		**92.03**	**92.31**	**92.17**	**93.50**	**93.57**	**93.53**

Table A.1: Continued

1	FIN-PROD	83.76	80.33	82.01	83.88	83.20	83.54
2		73.91	76.88	75.37	81.35	78.99	80.10
3		78.89	83.07	80.93	82.83	86.77	84.75
4		79.60	76.56	78.05	82.35	87.08	84.65
5		76.47	81.64	78.97	80.00	85.02	82.44
Average		**78.53**	**79.70**	**79.07**	**82.08**	**84.19**	**83.10**

1	GATHERING	71.21	75.81	73.44	76.67	74.19	75.41
2		65.74	80.68	72.45	78.57	87.50	82.80
3		67.01	74.71	70.65	81.11	83.91	82.49
4		65.56	73.75	69.41	64.95	78.75	71.19
5		80.60	76.06	78.26	80.82	83.10	81.94
Average		**70.02**	**76.20**	**72.84**	**76.42**	**81.49**	**78.77**

Table A.1: Continued

1	INDEX	87.89	92.78	90.27	93.99	95.56	94.77
2		87.61	91.67	89.59	94.57	96.76	95.65
3		89.81	87.00	88.38	90.28	87.44	88.84
4		87.37	92.02	89.64	93.23	95.21	94.21
5		83.78	89.08	86.35	81.91	88.51	85.08
Average		**87.29**	**72.87**	**88.85**	**90.80**	**92.70**	**91.71**

1	LOCATION	94.30	95.17	94.73	96.49	96.68	96.59
2		95.37	95.57	95.47	96.53	96.24	96.39
3		94.86	96.63	95.74	96.63	97.19	96.91
4		94.89	95.10	95.00	96.46	96.03	96.24
5		95.97	96.24	96.10	96.09	96.36	96.23
Average		**95.08**	**95.74**	**95.41**	**96.44**	**96.50**	**96.47**

Table A.1: Continued

1	LOTTERY	0.00	0.00	0.00	100.00	100.00	100.00
2		100.00	42.86	60.00	100.00	100.00	100.00
3		0.00	0.00	0.00	100.00	80.00	88.89
4		66.67	28.57	40.00	85.71	85.71	85.71
5		0.00	0.00	0.00	100.00	100.00	100.00
Average		**33.33**	**14.29**	**20.00**	**97.142**	**93.142**	**94.92**

1	MONETARY-VALUE	97.01	92.02	94.45	97.30	92.92	95.06
2		94.66	94.92	94.79	94.60	96.16	95.37
3		97.60	95.58	96.58	97.46	95.30	96.37
4		93.03	96.06	94.52	97.40	96.60	97.00
5		98.33	96.50	97.41	98.22	96.73	97.47
Average		**96.13**	**95.02**	**95.55**	**97.00**	**95.54**	**96.25**

Table A.1: Continued

1	ORG	90.55	92.03	91.29	93.78	93.87	93.83
2		90.92	92.54	91.73	93.84	94.65	94.24
3		91.69	92.48	92.08	93.62	94.82	94.22
4		91.45	93.56	92.49	94.44	95.27	94.85
5		91.83	93.56	92.68	93.30	94.66	93.97
Average		**91.29**	**92.83**	**92.05**	**93.80**	**94.65**	**94.22**

1	PERCENT	99.05	98.01	98.53	98.84	98.43	98.64
2		96.74	98.00	97.36	97.21	97.80	97.50
3		98.18	97.47	97.82	98.01	98.10	98.06
4		98.27	97.24	97.75	97.91	98.00	97.95
5		97.58	97.86	97.72	98.54	98.74	98.64
Average		**97.96**	**97.72**	**97.84**	**98.10**	**98.21**	**98.16**

Table A.1: Continued

1	PERSON	97.68	98.35	98.02	98.71	99.05	98.88
2		97.69	97.44	97.57	98.81	98.98	98.89
3		97.24	96.49	96.87	98.89	98.80	98.84
4		96.47	97.13	96.80	97.83	98.90	98.36
5		98.34	98.34	98.34	99.04	99.39	99.21
Average		**97.48**	**97.55**	**97.52**	**98.66**	**99.02**	**98.84**

1	TAX	93.02	90.91	91.95	95.45	95.45	95.45
2		89.19	94.29	91.67	91.89	97.14	94.44
3		91.43	91.43	91.43	91.18	88.57	89.86
4		94.29	86.84	90.41	100.00	97.37	98.67
5		100.00	94.64	97.25	100.00	100.00	100.00
Average		**93.59**	**91.62**	**92.54**	**95.70**	**95.71**	**95.68**

Table A.1: Continued

1	TIME	44.44	44.44	44.44	66.67	44.44	53.33
2		66.67	72.73	69.57	53.33	72.73	61.54
3		73.33	84.62	78.57	92.31	92.31	92.31
4		87.50	100.00	93.33	77.78	100.00	87.50
5		45.45	50.00	47.62	66.67	80.00	72.73
Average		**63.48**	**70.36**	**66.71**	**71.35**	**77.90**	**73.48**

Table A.2: Scores of DistilBERT Models for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

		DistilBERT Base Multilingual Cased			**DistilBERT Base Turkish Cased**		
Fold	**Named Entity**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
1	BUSINESS-SECTOR	73.23	76.11	74.64	75.57	76.64	76.10
2		73.47	76.13	74.77	74.30	78.38	76.29
3		74.42	79.54	76.89	75.86	77.45	76.64
4		73.91	77.75	75.79	75.00	77.65	76.30
5		72.12	73.36	72.74	74.57	74.57	74.57
Average		**73.43**	**76.58**	**74.97**	**75.06**	**76.94**	**75.98**
1	COMMODITY	89.02	88.44	88.73	89.79	91.69	90.73
2		88.73	89.68	89.20	87.52	91.10	89.28
3		88.30	90.69	89.48	90.08	93.74	91.87
4		88.32	89.61	88.96	88.92	92.53	90.69
5		86.45	91.72	89.01	88.66	93.04	90.80
Average		**88.16**	**90.03**	**89.08**	**88.99**	**92.42**	**90.67**

Table A.2: Continued

1	CURRENCY	97.83	96.55	97.19	98.56	97.50	98.03
2		99.27	97.27	98.26	98.68	97.75	98.21
3		97.86	97.86	97.86	98.20	98.20	98.20
4		98.52	98.19	98.36	98.75	98.19	98.47
5		98.54	98.13	98.34	98.54	97.82	98.18
Average		**98.40**	**97.60**	**98.00**	**98.55**	**97.89**	**98.22**

1	DATE	90.89	91.35	91.12	91.30	92.11	91.70
2		92.97	92.84	92.90	92.98	93.11	93.05
3		93.49	94.70	94.09	91.67	93.56	92.60
4		90.07	90.57	90.32	89.93	90.43	90.18
5		90.57	91.42	90.99	90.98	91.96	91.47
Average		**91.60**	**92.18**	**91.88**	**91.37**	**92.23**	**91.80**

Table A.2: Continued

1	FIN-PROD	78.23	79.51	78.86	77.05	77.05	77.05
2		75.90	74.37	75.13	71.29	72.36	71.82
3		75.37	80.95	78.06	76.00	80.42	78.15
4		77.73	81.82	79.72	71.55	79.43	75.28
5		73.87	79.23	76.46	69.86	73.91	71.83
Average		**76.22**	**79.18**	**77.65**	**73.15**	**76.63**	**74.83**

1	GATHERING	75.86	70.97	73.33	65.67	70.97	68.22
2		54.21	65.91	59.49	63.46	75.00	68.75
3		70.00	72.41	71.19	56.48	70.11	62.56
4		61.80	68.75	65.09	62.22	70.00	65.88
5		71.43	77.46	74.32	64.63	74.65	69.28
Average		**66.66**	**71.10**	**68.68**	**62.49**	**72.15**	**66.94**

Table A.2: Continued

1	INDEX	91.35	93.89	92.60	90.53	95.56	92.97
2		86.34	90.74	89.49	88.74	91.20	89.95
3		89.45	87.44	88.44	84.07	85.20	84.63
4		87.82	92.02	89.87	91.33	95.21	93.23
5		81.48	88.51	84.85	77.89	85.06	81.32
Average		**87.29**	**90.52**	**88.85**	**86.51**	**90.45**	**88.42**

1	LOCATION	94.06	94.75	94.40	94.65	94.86	94.76
2		94.75	95.11	94.93	94.57	94.85	94.71
3		94.31	95.72	95.01	94.87	95.74	95.30
4		94.65	94.62	94.64	94.48	95.21	94.84
5		94.16	95.23	94.69	94.18	95.39	94.78
Average		**94.39**	**95.09**	**94.73**	**94.55**	**95.21**	**94.88**

Table A.2: Continued

1	LOTTERY	50.00	20.00	28.57	50.00	20.00	28.57
2		100.00	28.57	44.44	100.00	42.86	60.00
3		0.00	0.00	0.00	0.00	0.00	0.00
4		100.00	28.57	44.44	100.00	28.57	44.44
5		0.00	0.00	0.00	0.00	0.00	0.00
Average		**50.00**	**15.43**	**23.49**	**50.00**	**18.29**	**26.60**

1	MONETARY-VALUE	96.69	90.35	93.41	95.76	90.09	92.84
2		93.66	95.20	94.42	95.59	95.20	95.40
3		96.33	94.20	95.25	96.04	93.78	94.90
4		93.60	95.38	94.48	95.19	96.74	95.96
5		95.43	95.09	95.26	96.78	94.74	95.75
Average		**95.14**	**94.04**	**94.56**	**95.87**	**94.11**	**94.97**

Table A.2: Continued

1	ORG	89.11	90.25	89.67	89.20	91.48	90.32
2		88.89	90.87	89.92	88.41	91.90	90.12
3		89.53	91.73	90.62	89.10	92.00	90.52
4		91.40	92.96	92.17	89.68	91.75	90.70
5		91.12	91.93	91.53	90.28	91.42	90.85
Average		**90.03**	**91.55**	**90.78**	**89.33**	**91.71**	**90.50**

1	PERCENT	99.16	98.33	98.74	98.84	98.43	98.64
2		97.42	98.10	97.76	96.35	97.90	97.12
3		98.27	97.56	97.91	98.09	97.47	97.78
4		98.18	97.52	97.85	98.17	97.14	97.65
5		98.05	97.76	97.91	97.47	97.57	97.52
Average		**98.22**	**97.85**	**98.03**	**97.78**	**97.70**	**97.74**

Table A.2: Continued

1	PERSON	97.32	97.49	97.40	97.24	97.58	97.41
2		96.84	96.76	96.80	97.62	97.87	97.74
3		91.28	95.81	93.49	97.60	97.35	97.47
4		96.34	95.77	96.06	96.55	97.13	96.84
5		97.38	97.64	97.51	97.47	97.73	97.60
Average		**95.83**	**96.69**	**96.25**	**97.30**	**97.53**	**97.41**

1	TAX	83.33	90.91	86.96	82.00	93.18	87.23
2		94.44	97.14	95.77	89.47	97.14	93.15
3		88.89	91.43	90.14	85.71	85.71	85.71
4		97.30	94.74	96.00	94.87	97.37	96.10
5		89.66	92.86	91.23	92.59	89.29	90.91
Average		**90.72**	**93.42**	**92.02**	**88.93**	**92.54**	**90.62**

Table A.2: Continued

1	TIME	66.67	44.44	53.33	66.67	44.44	53.33
2		58.33	63.64	60.87	66.67	72.73	69.57
3		73.33	84.62	78.57	85.21	92.31	88.89
4		87.50	100.00	93.33	77.78	100.00	87.50
5		40.00	40.00	40.00	77.78	70.00	73.68
Average		**65.17**	**66.54**	**65.22**	**74.92**	**75.90**	**74.59**

Table A.3: Scores of XLM-RoBERTa Base Multilingual and ELECTRA Base Turkish Cased Discriminator Models for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

		XLM-RoBERTa Base Multilingual			**ELECTRA Base Turkish Cased**		
Fold	**Named Entity**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
1	BUSINESS-SECTOR	77.92	76.96	78.93	77.48	81.07	79.23
2		76.35	78.96	77.63	77.94	82.29	80.06
3		78.00	81.74	79.82	79.69	83.03	81.33
4		76.56	81.50	78.95	75.78	83.58	79.49
5		74.36	76.90	75.61	77.63	79.03	78.33
Average		**76.64**	**79.81**	**78.19**	**77.70**	**81.80**	**78.69**
1	COMMODITY	90.92	93.00	91.95	92.13	93.49	92.81
2		89.51	91.10	90.30	90.83	93.42	92.11
3		90.55	94.08	92.28	93.85	95.60	94.72
4		91.67	92.86	92.26	92.22	94.32	93.26
5		88.53	94.72	91.52	89.77	94.90	92.26
Average		**90.24**	**93.15**	**91.66**	**91.76**	**94.35**	**93.03**

Table A.3: Continued

1	CURRENCY	98.69	98.45	98.57	99.05	98.69	98.87
2		99.40	97.86	98.63	99.16	97.98	98.57
3		98.10	98.65	98.37	98.65	99.10	98.88
4		98.86	98.30	98.58	98.98	98.76	98.87
5		98.75	98.65	98.70	99.17	98.86	99.01
Average		**98.76**	**98.38**	**98.57**	**99.00**	**98.68**	**98.84**

1	DATE	91.63	91.88	91.81	91.14	91.60	91.37
2		92.73	93.11	92.92	94.20	93.94	94.07
3		93.67	93.55	93.61	94.61	94.97	94.79
4		91.84	92.09	91.97	90.85	90.85	90.85
5		91.34	91.96	91.65	93.15	93.03	93.09
Average		**92.24**	**92.54**	**92.39**	**92.79**	**92.88**	**92.83**

Table A.3: Continued

1	FIN-PROD	84.55	80.74	82.60	84.00	86.07	85.02
2		79.96	73.87	75.38	71.82	79.40	75.42
3		73.02	83.07	77.72	80.49	87.30	83.76
4		78.90	82.30	80.56	82.35	87.08	84.65
5		82.52	82.13	82.32	78.70	87.44	82.84
Average		**79.19**	**80.42**	**79.72**	**79.47**	**85.46**	**82.34**

1	GATHERING	82.26	82.26	82.26	81.97	80.65	81.30
2		66.36	82.95	73.74	76.24	87.50	81.48
3		82.42	86.21	84.27	81.32	85.06	83.15
4		63.74	72.50	67.84	68.82	80.00	73.99
5		83.10	83.10	83.10	83.82	80.28	82.01
Average		**75.58**	**81.40**	**78.24**	**78.43**	**82.70**	**80.39**

Table A.3: Continued

1	INDEX	92.93	95.00	93.96	90.37	93.89	92.10
2		90.62	93.98	92.27	92.38	95.37	93.85
3		88.64	87.44	88.04	88.18	87.00	87.58
4		87.56	93.62	90.49	91.84	95.74	93.75
5		86.11	89.08	87.57	78.06	87.93	82.70
Average		**89.17**	**91.82**	**90.47**	**88.17**	**91.99**	**90.00**

1	LOCATION	95.27	95.64	95.45	95.47	95.91	95.69
2		94.45	94.85	94.65	96.55	95.61	96.08
3		96.13	96.68	96.41	96.77	96.84	96.81
4		94.93	94.43	94.68	96.10	96.10	96.10
5		95.30	95.88	95.59	95.62	96.16	95.89
Average		**95.22**	**95.50**	**95.36**	**96.10**	**96.12**	**96.11**

Table A.3: Continued

1	LOTTERY	0.00	0.00	0.00	0.00	0.00	0.00
2		100.00	28.57	44.44	0.00	0.00	0.00
3		0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.00	0.00	0.00	0.00	0.00
5		0.00	0.00	0.00	0.00	0.00	0.00
Average		**20.00**	**5.71**	**8.89**	**0.00**	**0.00**	**0.00**

1	MONETARY-VALUE	95.30	91.38	93.30	96.46	91.25	93.78
2		94.17	95.34	94.75	93.56	95.61	94.57
3		97.03	94.61	95.80	97.60	95.30	96.44
4		94.64	95.92	95.28	95.32	96.88	96.09
5		98.21	95.91	97.04	98.22	96.73	97.47
Average		**95.87**	**94.63**	**95.23**	**96.23**	**95.15**	**95.67**

Table A.3: Continued

1	ORG	92.43	92.90	92.67	93.12	93.26	93.19
2		92.91	93.52	93.21	92.30	93.95	93.12
3		92.63	93.22	92.93	92.99	94.08	93.53
4		92.90	94.22	93.56	93.95	94.83	94.39
5		93.83	94.03	93.93	92.56	94.45	93.50
Average		**92.94**	**93.58**	**93.26**	**92.98**	**94.11**	**93.55**

1	PERCENT	99.05	98.12	98.58	98.85	98.54	98.69
2		98.00	98.00	98.00	97.80	98.01	97.95
3		98.36	97.74	98.05	98.19	98.37	98.28
4		97.98	97.05	97.51	97.80	97.52	97.66
5		96.62	97.47	97.05	97.48	97.76	97.62
Average		**98.00**	**97.68**	**97.84**	**98.02**	**98.06**	**98.04**

Table A.3: Continued

1	PERSON	97.43	98.53	97.98	98.96	99.31	99.14
2		98.04	98.04	98.04	98.72	98.55	98.63
3		98.29	98.55	98.42	98.12	98.46	98.29
4		97.97	98.06	98.01	98.40	98.56	98.48
5		98.35	98.86	98.61	98.95	98.69	98.82
Average		**98.02**	**98.41**	**98.21**	**98.63**	**98.71**	**98.67**

1	TAX	93.18	93.18	93.18	87.76	97.73	92.47
2		91.89	97.14	94.44	82.93	97.14	89.47
3		88.57	88.57	88.57	86.11	88.57	87.32
4		95.00	100.00	97.44	94.87	97.37	96.10
5		96.43	96.43	96.43	98.25	100.00	99.12
Average		**93.01**	**95.06**	**94.01**	**89.98**	**96.16**	**92.90**

Table A.3: Continued

1	TIME	57.14	44.44	50.00	66.67	44.44	53.33
2		53.33	72.73	61.54	66.67	72.73	69.57
3		92.31	92.31	92.31	80.00	92.31	85.71
4		77.78	100.00	87.50	77.78	100.00	87.50
5		60.00	60.00	60.00	58.33	70.00	63.64
Average		**68.11**	**73.90**	**70.27**	**69.89**	**75.90**	**71.95**

Table A.4: Scores of ConvBERT Base Turkish Cased and mDeBERTa Base Multilingual Models for 5-Fold Cross Validation Experiments per Named Entity Type with FINTURK500K-PARENTS Dataset Annotated with BIO Schema

		ConvBERT Base Turkish Cased			**mDeBERTa Base Multilingual**		
Fold	**Named Entity**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
1	BUSINESS-SECTOR	78.61	80.66	79.62	76.75	80.46	78.56
2		78.49	82.49	80.44	76.87	82.58	79.62
3		78.27	84.83	81.42	78.01	83.53	80.67
4		79.86	82.85	81.33	74.90	81.91	78.25
5		77.09	79.79	78.41	76.29	78.69	77.47
Average		**78.46**	**82.12**	**80.24**	**76.56**	**81.43**	**78.91**
1	COMMODITY	95.59	95.60	94.07	93.88	92.51	93.19
2		92.51	92.35	92.43	90.69	93.59	92.12
3		93.38	95.43	94.39	91.00	94.08	92.51
4		92.69	96.75	94.68	89.91	93.99	91.90
5		92.11	93.82	92.96	90.64	95.31	92.92
Average		**92.66**	**94.79**	**93.71**	**91.22**	**93.90**	**92.53**

Table A.4: Continued

1	CURRENCY	99.28	98.45	98.87	99.16	98.34	98.75
2		99.40	98.10	98.75	99.40	97.75	98.56
3		98.65	99.10	98.88	98.32	99.10	98.71
4		98.76	98.87	98.81	99.09	98.42	98.75
5		99.37	98.76	99.06	99.17	98.86	99.01
Average		**99.09**	**98.66**	**98.87**	**99.03**	**98.49**	**98.76**

1	DATE	92.13	92.37	92.25	92.30	93.00	92.65
2		93.39	93.39	93.39	93.39	93.39	93.39
3		94.85	94.85	94.85	94.45	94.57	94.51
4		91.81	91.68	91.74	90.46	90.71	90.58
5		93.57	93.57	93.57	92.63	92.63	92.63
Average		93.15	93.17	93.16	92.65	92.86	92.75

Table A.4: Continued

1	FIN-PROD	85.89	84.84	85.36	80.86	84.84	82.80
2		75.86	77.39	76.62	72.99	77.39	75.12
3		82.84	89.42	86.01	78.00	82.54	80.21
4		79.48	87.08	83.11	78.83	83.73	81.21
5		79.56	86.47	82.87	83.92	80.68	82.27
Average		**80.73**	**85.04**	**82.79**	**78.92**	**81.84**	**80.32**

1	GATHERING	83.87	83.87	83.87	75.38	79.03	77.17
2		79.59	88.64	83.87	76.29	84.09	80.00
3		79.79	86.21	82.87	79.78	81.61	80.68
4		73.26	78.75	75.90	69.05	72.50	70.73
5		85.71	84.51	85.11	81.69	81.69	81.69
Average		**80.44**	**84.40**	**82.32**	**76.44**	**79.78**	**78.05**

Table A.4: Continued

1	INDEX	93.48	95.56	94.51	92.97	95.56	94.25
2		93.78	97.69	95.69	93.18	94.91	94.04
3		88.69	87.89	88.29	89.14	88.34	88.74
4		93.26	95.74	94.49	90.26	93.62	91.91
5		87.64	89.66	88.64	85.25	89.66	87.39
Average		**91.37**	**93.31**	**92.32**	**90.16**	**92.42**	**91.27**

1	LOCATION	96.24	96.80	96.52	96.14	96.06	96.10
2		96.16	96.24	96.20	96.30	96.58	96.44
3		95.79	96.84	96.31	96.09	96.67	96.38
4		96.33	95.55	95.94	95.87	95.70	95.78
5		96.11	95.96	96.04	96.24	96.20	96.22
Average		**96.13**	**96.28**	**96.20**	**96.13**	**96.24**	**96.18**

Table A.4: Continued

1	LOTTERY	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.00	0.00	0.00	0.00	0.00
3		0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.00	0.00	0.00	0.00	0.00
5		0.00	0.00	0.00	0.00	0.00	0.00
Average		**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

1	MONETARY-VALUE	96.77	92.66	94.67	96.35	91.63	93.93
2		97.61	95.20	96.39	95.21	95.47	95.34
3		97.59	94.89	96.22	97.74	95.72	96.72
4		95.97	97.01	96.49	94.69	96.88	95.77
5		97.41	96.85	97.13	97.88	97.20	97.54
Average		**97.07**	**95.32**	**96.18**	**96.37**	**95.38**	**95.86**

Table A.4: Continued

1	ORG	94.67	93.11	92.80	92.92	93.77	93.34
2		93.63	94.49	94.06	93.73	94.54	94.14
3		93.48	94.13	93.80	93.17	93.92	93.54
4		94.39	95.38	94.88	93.07	94.66	93.86
5		94.14	95.03	94.58	94.05	95.18	94.61
Average		**94.06**	**94.43**	**94.02**	**93.39**	**94.41**	**93.90**

1	PERCENT	98.74	98.01	98.37	98.95	98.74	98.85
2		97.80	97.90	97.85	97.31	97.60	97.45
3		98.46	98.28	98.37	98.28	98.10	98.19
4		97.25	97.71	97.48	97.99	97.62	97.81
5		97.88	98.74	98.31	98.07	98.64	98.35
Average		**98.03**	**98.13**	**98.08**	**98.12**	**98.14**	**98.13**

Table A.4: Continued

1	PERSON	98.70	98.79	98.75	98.27	98.53	98.40
2		98.72	98.72	98.72	98.89	98.46	98.67
3		98.64	99.40	99.02	97.88	98.63	98.25
4		98.82	99.07	98.94	98.07	98.82	98.44
5		99.21	99.30	99.26	99.21	99.39	99.30
Average		**98.82**	**99.06**	**98.94**	**98.46**	**98.77**	**98.61**

1	TAX	87.50	95.45	91.30	95.45	95.45	95.45
2		80.95	97.14	88.31	87.18	97.14	91.89
3		88.57	88.57	88.57	96.88	88.57	92.54
4		94.87	97.37	96.10	90.48	100.00	95.00
5		91.67	98.21	94.83	98.21	98.21	98.21
Average		**88.71**	**95.35**	**91.82**	**93.64**	**95.87**	**94.62**

Table A.4: Continued

1	TIME	66.67	44.44	53.33	71.43	55.56	62.50
2		81.82	81.82	81.82	53.85	63.64	58.33
3		92.31	92.31	92.31	92.31	92.31	92.31
4		77.78	100.00	87.50	77.78	100.00	87.50
5		70.00	70.00	70.00	63.64	70.00	66.67
Average		**77.72**	**77.71**	**76.99**	**71.80**	**76.30**	**73.46**