Named Entity Recognition in Sports Forecasts

Konovalov Kirill Kherullah Ghadeer

National Research University Higher School of Economics, Moscow

Abstract Ghadeer

Extensive studies in machine learning (ML) and deep neural networks (NN) are made in **Named Entity Recognition (NER)** since it is a basic first step for many natural languages processing tasks. Despite many state-of-the-art advancements that were made in NER, there are still lots of difficulties related to NER that we will shed the light on especially when it comes to extracting named entities from a bilingual database in Russian and English. Different methods related to Knowledge-based NER systems are also explored in extracting names of teams and leagues from sports forecasts. We will examine direct match, fuzzy search, datasets dictionaries, custom lists, and aliases dictionaries as well as vector models (W2v and FastText). We demonstrate that the deep learning technique of FastText proved to be the most efficient in NER for both English and Russian parts of the dataset.

1 Introduction

Kirill

The volume of data is rapidly increasing. Therefore, there is increasingly more demand for text mining tools for extracting information from all kinds of different text sources. The first step in information extraction is to detect the entities in the text. (Ghadeer)**Named entity recognition** means finding spans of text that constitute proper names (a person, a location, an organization even extends to include not exact entities per se, such as dates, times, and other kinds of temporal expressions, and even numerical expressions like prices) and then classifying the type of the entity (Jurafsky et al, 2019).

Named Entity Recognition is one of the most popular Natural Language Processing tasks. The problem itself was formulated over 20 years ago at the Message Understanding Conference (MUC, 1995; Sundheim, 1995) and nowadays it is widely used as a key component in systems for relation extraction, information retrieval, question answering, co-reference resolution, topic modeling, linking text to information in structured knowledge sources, etc.

Although researchers pay a lot of attention to business, biomedical, legal and other similar spheres, the sphere of automatic sports information extraction is underdeveloped, especially in the Russian language. In this research, we explore the

problem of named entity recognition in sports forecasts texts in both Russian and English.

We find named entities of team names and leagues in the dataset. Moreover, we look into the complexity of the task describing and classifying the problems which occurred during the process.

The non-standard structure of a sports forecast is one of the difficulties of this work. Such texts are written by sports fans and contain a lot of data that create additional noise and obstruct further work with the text. Besides, the difference in English and Russian sets another challenge as well.

Since team names and league names represent closed lists which can be covered by classical rule-based methods, we are more focused on them in our task.

(Ghadeer) In the following sections, we begin by giving a quick overview on our data (section 2). Then, we give a detailed study of all the aforementioned methods and approaches we used throughout our study (section 3). After that, we analyze the results obtained in our experiment and compare them to see which method performed best (section 4). Finally, we conclude by drawing conclusions from our research (section 5).

2 Data

Kirill

In our project, we are using a dataset in tsv format with forecasts written in Russian and English about various sports events. The dataset contains 870 362 texts with prognoses which have been parsed from sports fan websites and forums. Besides predictions it has information about team names, leagues, bookmakers, names, dates, tips and odds of different matches. Our dataset consists of the following columns:

- **Result.tsv:** the forecast and sometimes observations.
- **Odd**: the coefficient of the bet.
- **Tip_name:** the betting text in the forecast.
- **Bookmaker_name:** the company which accepts the bet.
- **Match_date_time:** the time of the match the forecast is concerned with. This is mostly missing in the forecasts.
- **League_name_in_russian:** the name of the sports league
- **Team_of_season_1_name_in_russian:** the name of the first competing team
- **Team_of_season_2_name_in_russian:** the name of the second competing team.

The first task was to filter out the repetitions in the data. With that done, our final version of the dataset has 702 908 unique forecasts. In addition to that, we

investigated the specificities of our data which was essential to better understanding it and finding the most efficient methods to be used. The details are the following:

- Total forecasts - 870 362
- Unique forecasts - 702 908
- Unique Russian forecasts - 595 931
- Unique English forecasts - 106 940
- Average length of a forecast - 52
- Median length of a forecast - 42

The forecasts of the matches are completely different and do not possess any common specific structure. Most of them contain information about the future match and the name of at least one team. Unfortunately, the league name is very often not present in the text of the forecast.

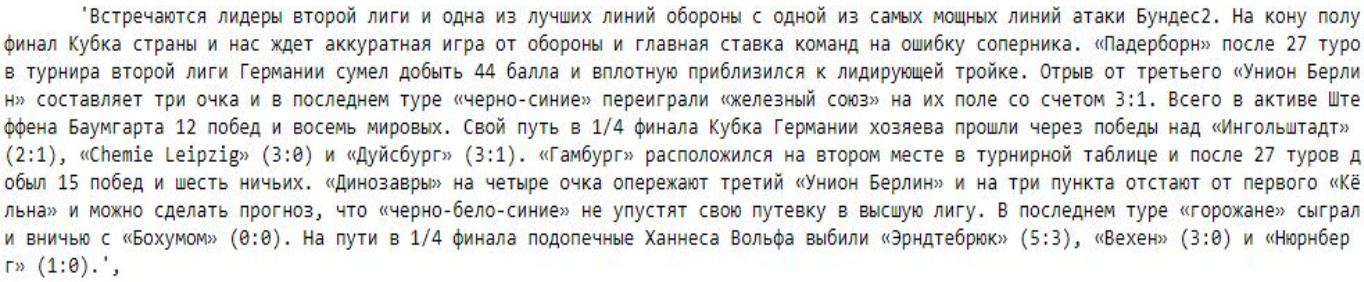


Fig. 1: Forecast from the dataset in Russian

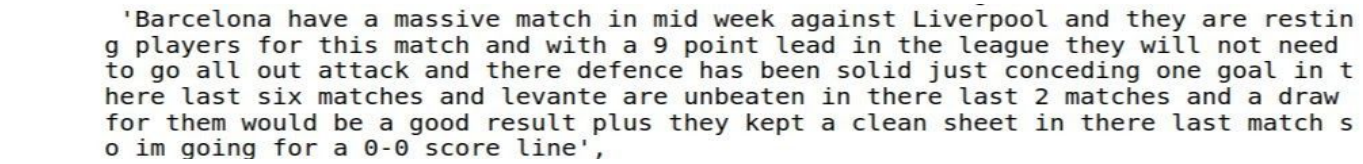


Fig. 2: Forecast from the dataset in English

All the data is available on Github.

3 Methodology

Kirill/ Ghadeer

There are a variety of solutions capable of extracting named entities even from very hard and multi-domain data. Such NER systems can be divided into four groups (Vikas Yadav at al, 2019): a) Knowledge-based systems (They do not require annotated training data as they rely on lexicon resources and domain specific knowledge. These work well when the lexicon is exhaustive.), b) Unsupervised and bootstrapped systems (Such models use seeds to discover text having potential named entities, detect noun phrases and filter any with low IDF values, and then

feed the filtered list to a classifier.), c) Feature-engineered supervised systems (These models learn to make predictions by training on example inputs and their expected outputs, and can be used to replace human curated rules.), d) Feature-inferring neural network systems (They can be broadly classified into categories depending upon their representation of the words in a sentence. For example, representations may be based on words, characters, other sub-word units or any combination of these.).

As was mentioned before, we apply the first group of methods and build knowledge-based systems which are able to recognize the specific entities needed for our task. This list of our methods includes direct match, fuzzy search,dataset lists, custom lists and dictionaries, dictionaries of aliases and vector models.

(Ghadeer) We started our research by getting familiarized with the dataset since it is essential for preprocessing the text and deciding the methods to be used for extracting named entities. The dataset was separated into two sub-datasets (one for Russian texts and the other for English). The same methods were performed independently on each of the sub-datasets. Then, at a later stage, results of different methods are compared to see the quality of each of the methods and what methods performed best, in each of the sub-datasets as a first step, and in the whole dataset as a second step.

The following preparation stages were conducted before the start of using each of the methods:

- Cleaning both parts of the dataset from noise (URLS), missing values (NaNs), duplicate values and other data abnormalities.
- Translating all teams and leagues names under the columns (League_name_in_russian, team_of_season_1_name_in_russian, team_of_season_2_name_in_russian) into english to be used in each of the methods under concern. This technique improved the results of the English sub-dataset significantly in all the methods used.

3.1.1 Direct match

Ghadeer

"Direct matching is principally defined by dictionary-based approaches for identifying concepts" (Gelbukh, 2011). It is the technique of searching for exact matches.

After the aforementioned preparation stages were conducted, we started the process of searching for exact matches for team 1, team 2 and leagues in each of the sub-datasets separately. We got good results from direct matching teams' names; however, it was not the case for leagues' names as shown in Table 1 for results of the Russian sub-data set and Table 2 for the English one. That is mainly because leagues names are almost not present in the texts of both Russian and

English forecasts as shown in mismatch analysis in the section entitled (Classification of the mismatches).

Labels	Direct match
team 1/команда 1	0.47 (47 %)
team 2/команда 2	0.44 (44 %)
league/лига	0.0497 (4.97 %)

Table 1: Direct Match results for Russian dataset

Labels	Direct match
team 1/команда 1	0.43 (43 %)
team 2/команда 2	0.41 (41 %)
league/лига	0.0397 (3.97 %)

Table 2: Direct Match results for English dataset

As shown in the tables, the direct match results in Russian are a little bit better than the English that is mainly due to having a larger sub-dataset in Russian. The direct match results are set as our baseline.

3.1.2 Classification of the mismatches

Kirill/ Ghadeer

Having considered the results of the direct match approach, we decided to classify the mismatches into categories. We extracted the texts of 300 random predictions in Russian and in English (for each) with mismatches in all three columns (team 1, team 2, league). Based on them, we built two tables for the typical problems (Table 3 for the mismatches in the Russian sub-dataset and Table 4 for the mismatches in the English one).

Russian sub-dataset:

Type of mismatch	Examples	Number of occurrences	%
1) Incomplete name of the league/team	Реал Мадрид - Реал, Атлетико Мадрид - Атлетико	43	14%
2) Different word form	Международный Кубок чемпионов - Международном кубке чемпионов, Премьер-лига - Премьер-лиги	28	9%

3) Abbreviations	Международный Кубок чемпионов - МКЧ, Атлетико Мадрид - АТМ, Лига Европы - ЛЕ, Премьер-лига - АПЛ	6	2%
4) Quotation marks	Динамо Киев - «Динамо» Киев	1	<1%
5) Combinations: incomplete names and/or different word forms and/or quotation marks etc.	Олимпик Донецк - "Олимпику", Женский чемпионат - чемпионата WNBA	48	16%
6) Aliases and other words	Бавария - мюнхенцы, Реал Мадрид - мадридцы, Примера - чемпионат Испании, Реал Мадрид - сливочные, Примейра-лига - чемпионат Португалии	79	26%
7) No data (team 1, team 2, league) in the text	Челси Арсенал (Финал лиги Европы, и он получился английский, также и финал лч. Интересный матч будет, 0-0 при этом исключать полностью не приходится - 1%, но прежде всего акцент на тотал, и от каждой как минимум по одному мячу. Доп время также может произойти, но ставки на иное вообще. Первый тайм будут команды бить, выносить мячи, не всегда туда - куда нужно в лучшем случае. Как минимум 5 на двоих соорудят для тотала, и при этом каждая не меньше двух раз подаст.) Международные товарищеские матчи (Сборная Нигерии после чемпионата мира провела 6 матчей и ни разу не проиграла, уверенно обеспечив себе путевку на Кубок Африки. Победа в недавнем поединке со сборной Сейшельских островов (3:1) позволила команде сохранить первое место в группе. Подопечные Гернота Рора владели подавляющим преимуществом и создали немало моментов у чужих ворот, поэтому в их успехе не было сомнений. Стоит отметить, что состав нигерийцев нельзя считать оптимальным: нет в заявке Келети Ихеаначо, Виктора Мозеса, Уилфрида Ндиди и многих других из числа тех, кто выступал на мундиале в России.)	259	86%
8) Author’s changes in the word and/or mistakes	Сиэтл Шторм - Сиэтл Сторм, Коннектикут Санс - Коннектикут Сан, Локомотив - локоиив, Хаммарбю Норрчепинг - хамарбро норчепинг	11	4%
9) Different register	Ренн Кан - ренн кан	24	8%
10) Errors in the dataset	Predictions about several matches in one, incorrect teams/leagues, incorrect data	8	2.5%

11) Different language	Мое мнение об это матче اللاعب الأول ليس لاعب قوي ، ولكن لديه نتائج جيدة في البطولات التي تعقد على الرمال. نتائج اللاعب الثاني أكثر استقرارا اللاعب الأول ناجح جدا في المجموعة الأولى. اللاعب الأول بالتأكيد سيفوز بمجموعة واحدة اللاعب الثاني سيفوز بفضل خبرته	2	<1%
12) Spam texts	Ждем интересной игры. Ждем интересной игры. Ждем интересной игры. Ждем интересной игры. Ждем интересной игры. Ждем интересной игры. Ждем интересной игры. Интуиция, Интуиция , Интуиция ,Интуиция,Интуиция, Интуиция, Интуиция, Интуиция, Интуиция, Интуиция, Интуиция. Лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг лейпциг	3	1%

Table 3: Mismatch analysis for the Russian dataset

English sub-dataset

Ghadeer

Type of mismatch	Example	Number of occurrences	%
1) Incomplete name (team / league)	-Macclesfield Town (Macclesfield) -Coventry City (Coventry) -Doncaster Rovers (Doncaster) - Grimsby Town (Grimsby) -Manchester city (City) -Newcastle United (Newcastle) -Leicester City (Leicester) -Crawley Town (Crawley) -Norwich City (Norwich)- -Manchester United (United)	26	8.6%
2) Acronyms	- MK Dons (MKD) -New York city (NYC)	3	1%
3) Aliases	-Manchester United (The Blues) - Wolverhampton (Wolves) -Côte d'Ivoire (Ivory Coast)	3	1%
4) Different register	- New York city (New York City) - Genk (genk) -Ghana (ghana) -Argentina (argentina) - gold Cup (Gold Cup)	7	2.3%

5) Word change and/or spellings mistakes	-Gibraltar (Glbraltar) -Zenith (Zenit) -Portland Timberz (Portland Timbers) -DR Congo (D R Congo) -Morocco (Marocco) -Lillestrom SK (Lillestrom's)	10	3.3%
6) No data (team 1, team 2, league) in the text or the names are expressed indirectly	Coventry City (Both teams are at top shape before this game, each of them is still undefeated at League 1 after 6 played games - by 3 wins and 3 draws. I expect this game to ended in draw too.) Amiens (Both teams can potentially go down but a point may be good enough if Caen lose a difficult match away at Lyon.) Egypt (This is another matchup in which neither side really has the initiative to win. Splitting the points would see both sides go through and I expect exactly that to happen here.) DR Congo (Both teams are in poor form and they have drawn two in the last three meetings. Therefore, I think this will be a draw.) Qatar (This match might cause an upset as Afghan players are very much strong and they have done well in their previous matches) Ireland (I expect a tough physical match without many goals where set pieces will be important as will defensive discipline on both sides) Dallas The away team are much higher in the league and the head to heads are very tight. A draw looks like the value to follow here	199	66.3%
7) Russian written in latin/ English characters	Oba narmalnie komandi nobvsioje dumaiu chto vigraet csk oninumeiut v gastyax vesti sebya dastoino., Nonuj etotbdvainoi shans delaet bezopasnim stavku ya za to chto stavit ot banka bolshe vigrat chut menshe.. No chtobi bilo bi jeleznim. Poetamu savlu dvoinoi shans i budu nadeica na luchoe.. Ya toje chwlovek i kak vse xachu zarabotat nemnoshko babla vot i vsio	18	6%
8) Different languages	Zenith/ Chinese (如果你可以召唤力量，请拖我 我无法抑制紧迫感 你必须慢慢下降 然后给那些粘稠的钥匙加油 哄骗我，我的爱 并	3	1%

	旋转我的螺旋桨 这是一个必要的邪恶 没有紧急情况的原因 借用白头鹰的喙 哦，瞬间的协同作用 哄骗我，我的爱 沉入明天 哄骗我，我的爱 并旋转我的螺旋桨 GreatGFFFyfffFmfjfjrueyekeifjFjdjFjdjfjrueuekeiej) Wolfsburg / Turkish (istenildiği kadar sözcük türetilebilmesine olanak sağlamaktadır.[14] Türkiye Türkçesi bu yönünden dolayı diğer Türk dilleriyle ortak ya da ayrık bulunan onlarca eke sahiptir.[15] Türkçe geniş kullanımıyla birlikte zengin bir dil olmasının yanı sıra,[kaynak belirtilmeli] genel itibarıyla “özne-nesne-yüklem” biçimindeki cümle kuruluşuna sahiptir.)		
9) Errors in the dataset	-Napoli (X 2) - Tottenham Hotspur © __T_T_	4	1.3%
10) Spam texts	Ecuador (DF of of of the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the the) Blackburn Rovers (vipibettt vipibettt) Premier League (w. ry yi nvpn bcc mb zx it rru oiu mb cb et z cb tr uy nb tty rer uy vv. mnb cct tt rr. uu. ct bu cr vc byb vc vbu uu tr rt iuy nv xv nvn bu yu iy ibvc tr iy oi oib nv vc tt yui bv xr cy bu nu yt ut ui ib bct bc vc iy ty uu bc cx rt oi iub nn mo cr tt yy et iu iu bui nuu vb cvxv nui iu ryi ib iy nci oby vu bci nv bc bn ct yv tu uv uvt uy uv yi tyb buv vub ib biv)	10	3.3%
11) Combination of reasons	-UEFA Champions League / latin +missing data (Iuventus bate gol) - Blackburn Rovers (Blackburn’s) incomplete name + word change -Portland Timberz (Timbers) incomplete name + word change - New York Red Bulls (NY) incomplete name + acronym -Manchester city (The citizens) Alias + missing data	17	5.6%

Table 4: Mismatch analysis for the English dataset

All the reasons almost the same for English and Russian, except for two:

- Unlike Russian, we do not have mismatches in English due to quotations marks.
- 6% of the mismatches in English are due to having Russian text written in latin/ English characters.

The main type of mismatch for both English and Russian, as the tables show, is having missing data related to (team 1, team 2, league) in the text or the names are expressed indirectly or referred to differently.

The analysis helps not only understanding the sources and types of problems we have, but also in finding proper ways of improvement not only in direct match but also in almost all the other methods since such problems are not only restricted to direct match.

3.2 Fuzzy search

Ghadeer

Fuzzy search is also called "**Approximate String Matching"** and refers to the technique of searching for approximate matches rather than exact ones (Gusfield, 1997; Navarro, 2001). It is used in record linkage, spam filtering, and spell checking which in particular can be helpful in our dataset. **Levenshtein distance** is used to calculate the difference between two strings. Two attempts (Ngram and Nongram variants) were tried for most of the columns in concern in each of the Russian and English sub-datasets.

As for the Russian dataset, the results obtained from n-gram variants were better than Non-grams for the labels team 1, team 2, league (Table 5). The same situation goes as well for the English sub-dataset (Table 6). The Non-gram variant performed better for all the labels team 1, team 2, as well as league. The fuzz ratio in the Non-gram variant was mostly set somehow low in order to get similar results to the ones we got from the direct match in english. The results of Ngrams variant were even much worse. Hence, the non-gram variant would be adopted while analyzing the results of the English sub-dataset in Section 4.

Labels	fuzzy search
team 1/команда 1	-Non-ngram: 0.61 (61 %)
	-Ngram: 0.62 (62 %)
team 2/команда 2	-Non-ngram: 0.59 (59 %)
	-Ngram: 0.60 (60 %)
league/лига	-Ngram: 0.29 (29 %)

Table 5: Fuzzy Search results for Russian dataset

Labels	fuzzy search
team 1/команда 1	-Non-ngram: 0.43 (43 %)
team 2/команда 2	-Non-ngram: 0.41 (41 %)
	-Ngram: 0.0016 (0.16 %)
league/лига	-Non-ngram: 0.039 (3.97 %)
	-Ngram:0.006 (0.6 %)

Table 6: Fuzzy Search results for English dataset

As the tables 5 and 6 show, fuzzy search method performed better for Russian sub-dataset than the English one, that is mainly because the Russian part is larger.

3.3 Dictionaries with data from the dataset

Ghadeer

One of the approaches adopted was to extract all teams and leagues names from both English and russian parts of the dataset (separately) to create dataset dictionaries of teams names in english and another one for teams names in Russian. The same thing was done for league names (where two dictionaries were created; one in English and the other one in Russian). We ended up having a dictionary of team names in russian that contains 4442 unique entries and the leagues dictionary in Russian comprises 288 unique entries. Whereas the English teams dictionary has 1233 unique entries for both team 1 and team 2 and the league's dataset dictionary has 78 unique entries. Then, these dataset dictionaries were used with direct match in each of the subsets of our data. However, this technique did not give good results neither for Russian (Table 7) nor for English (Table 8), though it performed better with the Russian sub-dataset.

Labels	Dataset lists/ dictionaries + direct match
team 1/команда 1	0.2415 (24.15%)
team 2/команда 2	0.2419 (24.19%)
league/лига	0.0 (0%)

Table 7: Data dictionaries search results for Russian dataset

Labels	Dataset lists/ dictionaries + direct match
team 1/команда 1	0.0006 (0.06%)
team 2/команда 2	0.0005 (0.05%)
league/лига	0.0 (0%)

Table 8: Data dictionaries search results for English dataset

3.4 Custom dictionaries

Kirill

Another approach that we tested was finding named entities with the help of custom dictionaries of team names and leagues in Russian and English. The data for these dictionaries was parsed with the Beautiful Soup package from wikipedia and stored as separate python dictionaries. The resulting collections for teams in Russian contain 4210 unique entities whereas the leagues dictionary has 97 unique values. In the English dictionary, we have 286 entries for leagues and 10818 entries for teams. All teams and leagues are divided by country for convenience.

We used the same method as we did with the dataset dictionaries in order to find named entities.

Labels	Custom dictionaries
team 1/команда 1	0.1792 (17.9%)
team 2/команда 2	0.1814 (18.1%)
league/лига	0.0 (0%)

Table 9: Custom dictionaries search results for Russian dataset

Labels	Custom dictionaries
team 1/команда 1	0.0015 (0.15%)
team 2/команда 2	0.00095 (0.095%)
league/лига	0.0 (0%)

Table 10: Custom dictionaries search results for English dataset

As we can see from the results, in the Russian dataset the custom dictionaries were worse than previous approaches, achieving accuracy of 17.9% for team 1 and 18.1% for team 2. In the English dataset, we scored 0.15% for team 1 and 0.095% for team 2. We were not able to extract leagues with this method, which

is mainly connected with the absence of them in the forecasts. Also, the custom dictionary method does not give the opportunity to find the entities in various word forms.

3.5 Aliases dictionaries

Kirill

Official football club names are not always represented in the texts of sports forecasts in their original forms which creates a problem for recognizing them as named entities. Moreover, some famous teams have a number of different nicknames and aliases. For this reason we decided to create a dictionary of such terms. We parsed such websites as wikipedia, sports.ru, pikabu and others to build a python dictionary of aliases of teams in Russian where the keys are nicknames of a team and the values are lists of official names. Some teams have similar or the same nicknames so they share the same lists.

```
'жёлто-зелёные': ['анжи', 'кубань', 'нант'],
'крылышки': ['крылья советов'],
'перья': ['крылья советов'],
'крысы': ['крылья советов', 'нак', 'жерминаль', 'харельбекке'],
'самарцы': ['крылья советов'],
'волжане': ['крылья советов'],
'приморцы': ['луч-энергия'],
'дальневосточники': ['луч-энергия'],
'фонари': ['луч-энергия'],
'аквитанцы': ['бордо'],
'жирондийцы': ['бордо'],
```

Fig. 3: Part of the dictionary with team aliases for Russian

Having applied the match with the dictionary of aliases, we got a very good accuracy. Even without the application of other methods, the dictionary of aliases is able to discover a good number of named entities.

Labels	Aliases match
team 1/команда 1	0.10 (10%)
team 2/команда 2	0.08 (8%)

Table 11: Aliases dictionary search results for Russian dataset

We only applied this method for the Russian part of the dataset since the forecasts in the English part barely contain any aliases at all.

3.6 Combinations of methods

Kirill

In order to achieve the best results, we decided to use several methods together to see how well they perform in different combinations. Besides that based on our classification of the problems analysis, we filtered out the forecasts that were shorter than 30 symbols as such texts did not contain any useful information or were spam.

3.6.1 Direct match and aliases:

Firstly, we combined the direct match method with the aliases dictionary for the teams in the Russian language.

Labels	Direct match and aliases dictionary
team 1/команда 1	0.496 (49.6%)
team 2/команда 2	0.473 (47.3%)

Table 12: Direct match and aliases dictionary search results for Russian dataset

This group of methods along with the filtration of the dataset achieved 49.6 % accuracy for team 1 and 47.3 % for team 2. The scores are better than our baseline for these labels in Russian.

3.6.2 Fuzzy search and aliases:

Another combination that we applied in our research was fuzzy search and the dictionary of aliases.

Labels	Fuzzy search and aliases dictionary
team 1/команда 1	0.699 (69.9%)
team 2/команда 2	0.67 (67 %)

Table 13: Fuzzy search and aliases dictionary search results for Russian dataset

By using these methods, we got the accuracy of 69 % for team 1 and 67% for team 2 in the Russian dataset. These results beat the baseline and the results of the previous combination.

3.6.3 Custom dictionaries and fuzzy search:

The last combination of methods to be tested was custom dictionaries along with fuzzy search. However, we did not manage to apply it to the whole dataset neither in Russian nor in English due to its time-consuming process. The method uses custom dictionaries which contain 4210 unique teams and 97 leagues in Russian and 10818 teams and 286 leagues in English.

In order to reduce the time needed for the algorithm to search for approximate matches between the words of a forecast and our custom dictionaries, we subset our datasets so that they only had 10000 texts. Moreover, we decided not to check all the words of a forecast but only the ones in capital letters and in quotation marks.

Labels	Custom dictionaries and fuzzy search
team 1/команда 1	0.645 (64.5 %
team 2/команда 2	0.62 (62 %)
league/лига	0.164 (16.4 %)

Table 14: Custom dictionaries and fuzzy search results for Russian dataset

Labels	Custom dictionaries and fuzzy search
team 1/команда 1	0.74 (74%)
team 2/команда 2	0.71 (71%)
league/лига	0.08 (8%)

Table 15: Custom dictionaries and fuzzy search results for English dataset

The results of the method for both Russian and English dataset are quite high. It achieved accuracy of 64.5% for team 1, 62.5% for team 2 and 16.4% for leagues in the Russian part. In the English part the figures are 74% for team 1, 71% for team 2 and 8% for leagues.

3.7 Vector models

Ghadeer

Word embeddings or **word vectors** are learnet dense vector representations of words in which words sharing similar meanings also share similar representations. It is one of the deep learning techniques. The main advantage of word embeddings over traditional approach of one-hot representations is that similar

words have similar embeddings, while one hot vectors are orthogonal, thus, similarity cannot be computed (Collobert et al, 2011).

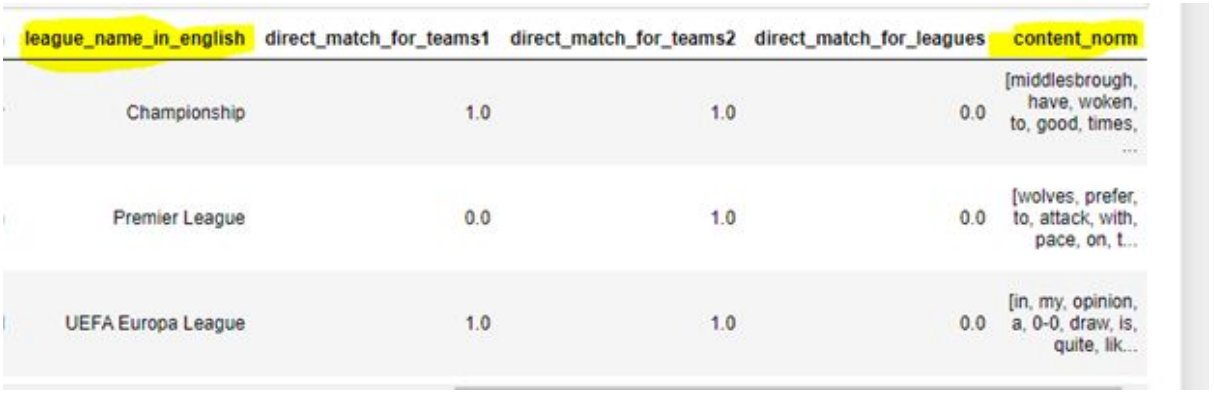
It was until 2013, Mikolov et. al proposed **Word2vec** approach which is much more efficient than the old different matrix factorization algorithms which were previously used to compute similarity. One of the advantages of Word2vec is that it can be trained on much larger datasets, hence positively affecting the quality of the embeddings.

Word2vec and its later introduced improvement **FastTex**t (Athiwaratkun et al., 2018) are widely used because they are much faster and easier to train than the state-of-the-art contextual embeddings from ELMO and BERT. FastText model is “fast, allowing to train models on large corpora quickly and allows us to compute word representations for words that did not appear in the training data” (Bojanowski et al, 2017).

Word embeddings are trained using unlabeled textual data. For the purpose of this research, embeddings are computed using W2V and FasText. Both models were trained for the Russian and English independently. We used four sets of vectors, two for each the English part (one Word2vec and one FastText) and two for the Russian part (one Word2vec and one FastText). These sets differ in the way they are created:

1. First set is English language set created by word2vec model
2. Second set is English language set created by fasttext model
3. Third set is Russian language set created by word2vec model
4. Fourth set is Russian language set created by fasttext model

W2vec and FastText models are simply used to first generate the embeddings of words (which are, for example for the English dataset, team name 1 in English, team name 2 in English, and league name is English) and then we calculate the embedding of column ‘content_norm’, which is actually derived from the column results. So after generating the embeddings, we have calculated the similarity that either the team1 or team2 or league name is present in the content norm column or not (Figure: 4):



league_name_in_english	direct_match_for_teams1	direct_match_for_teams2	direct_match_for_leagues	content_norm
Championship	1.0	1.0	0.0	[middlesbrough, have, woken, to, good, times, ...
Premier League	0.0	1.0	0.0	[wolves, prefer, to, attack, with, pace, on, t...
UEFA Europa League	1.0	1.0	0.0	[in, my, opinion, a, 0-0, draw, is, quite, lik...

Fig. 4: Part of the vector models from the English dataset

As for both the Russian (Table 16) and English (Table 17) parts of the dataset results, we got very good results in the FasText model for many reasons:

a) FastText uses subword pieces for inferring a word vector. So, in the FastText model, we get not only semantically similar words but also words that are similar in spelling.

b) FastText can also predict a vector for an unseen word if it consists of known-to-the-model subword pieces. It can be particularly useful when there are typos in the texts.

Labels	FastText	W2v
team 1/команда 1	0.71 (71%)	0.40 (40%)
team 2/команда 2	0.70 (70%)	0.38 (38%)
league/лига	0.03 (3%)	0.04 (4%)

Table 16: FasText and W2V models results in Russian dataset

Labels	FastText	W2v
team 1/команда 1	0.78 (78%)	0.48 (48%)
team 2/команда 2	0.75 (75%)	0.40 (40%)
league/лига	0.178 (17.8%)	0.022 (2.2%)

Table 17: FasText and W2V models results in English dataset

So, with the FastText model, we could solve many of the mismatch types such as spelling mistakes or change in the form of the word, different registers, and incomplete names.

FastText showed much better performance than the W2v model for both English and Russian. Besides, it proved to be very efficient especially for the English sub-dataset for which the results are even a little bit better than those of the Russian.

4 Results and analysis

Ghadeer

4.1 Comparison of the results:

We conduct a comprehensive comparative study that is divided into two parts: One is dedicated to comparing the different methods used in each of the individual sub-data sets, that is English and Russian (separately). The second focuses on comparing the results of all of the methods obtained in Russian sub-data set to the ones obtained in English sub-data set.

4.1.1 Comparison of methods (in each part of the dataset)

All the results we have from the different methods discussed in the previous sections are listed, from best to worst, for Russian (Table 18) and English (Table 19):

- **Russian dataset:**

This section is dedicated to analyzing the results of the ten different methods (individual and combined) used to extract the named entities (team1, team2, league):

- team1: Best accuracy is realized with FastText (71%), followed by fuzzy search with aliases (69.9%), then custom dictionaries with fuzzy search (64.5 %), after which comes fuzzy search/ n-gram (62 %), then direct match with aliases (49.6%), followed by direct match (47 %), W2v (40 %), then dataset lists with direct match (24.15%), custom lists with direct match (17.9%), and lastly aliases dictionaries (10%).

- team2: Best results are with FastText (70%) as well, followed by fuzzy search with aliases (67 %), then custom dictionaries with fuzzy search (62 %), after which comes fuzzy search/ n-gram (60 %), then direct match with aliases (47.3%), followed by direct match (44 %), then W2v (38%), dataset lists with direct match (24.19%), custom lists with direct match (18.1%), and lastly aliases dictionaries (8 %).

- league: The highest accuracy is achieved with fuzzy search/ Ngram (29 %), followed by custom dictionaries with fuzzy search (16.4 %), then direct match (4.97%), after which comes W2v (4%) and fastText (3%), and lastly both dataset dictionaries with direct match and custom dictionaries with direct match (0 %).

Labels	team 1/команда 1	team 2/команда 2	league/лига
FastText	**0.71 (71%**)	**0.70 (70%)**	0.03 (3%)
Fuzzy search and aliases	0.699 (69.9%)	0.67 (67 %)	-
Custom dictionaries and fuzzy search	0.645 (64.5 %)	0.62 (62 %)	0.164 (16.4 %)
Fuzzy search	-Non-ngram: 0.61 (61 %) -Ngram: 0.62 (62 %)	-Non-ngram: 0.59 (59 %) -Ngram: 0.60 (60 %)	Ngram: **0.29 (29 %)**
Direct match and aliases	0.496 (49.6%)	0.473 (47.3%)	-
Direct match	0.47 (47 %)	0.44 (44 %)	0.0497 (4.97 %)
W2v	0.40 (40%)	0.38 (38%)	0.04 (4%)
Dataset lists/ dictionaries + direct match	0.2415 (24.15%)	0.2419 (24.19%)	0.0 (0%)
Custom lists/ dictionaries + direct match	0.1792 (17.9%)	0.1814 (18.1%)	0.0 (0%)
Aliases dictionaries	0.10 (10 %)	0.08 (8 %)	-

Table 18: Results of all the methods used on the Russian part of the dataset

As we we can see:

- combined methods with Aliases realized high accuracy for the Russian sub-dataset, while the lowest accuracy was realized with aliases when not combined with any of the other methods.
- Fuzzy Search proved to achieve very good results for the Russian sub-dataset while used alone and even better when combined with other methods such as custom and aliases' dictionaries.
- Hence, Fuzzy Search when combined with aliases achieved very high accuracy. However, it could not out do FastText which performed the best for Team 1 (71%) and Team 2 (70%), but not for leagues. They are followed by custom dictionaries with fuzzy search, fuzzy search, recording the highest accuracy for leagues (29%) . After that comes direct match with aliases, direct match, W2v, data lists with direct match, custom lists with direct match, and finally aliases dictionaries. leagues were not extracted with aliases or any other methods combined with aliases and scored

0% , as expected due to their absence, in dataset dictionaries with direct match and custom lists with direct match.

- **English dataset:**

Now, we will closely examine the results achieved with the seven methods used to extract the following named entities from the English sub-data set:

- teams 1: Best accuracy is achieved with FastText (78%), followed by customs lists with fuzzy search (74%), then W2V model (48%), after which comes direct match as well as fuzzy search/ non-ngram (43%), then custom lists with direct match (0.15%), and lastly dataset lists with direct match (0.06%).
- teams 2: Best result is achieved with Fasttext (75%), after which comes custom dictionaries with fuzzy search (71%), followed by direct match as well as fuzzy search/ non-ngram (41%), then W2V model (40%), then custom lists with direct match (0.095%), and lastly dataset lists with direct match (0.05%).
- league: Highest accuracy is achieved with Fasttext (17.8%), then custom lists with fuzzy search (8%), followed by direct match as well as fuzzy search/ non-ngram (3.97%), then W2V model (2.2%), and lastly each of dataset lists with direct match and custom lists with direct match (0%) .

Labels	team 1/команда 1	team 2/команда 2	league/лига
FastText	**0.78 (78%)**	**0.75 (75%)**	**0.178 (17.8%)**
Custom lists/ dictionaries +Fuzzy Search	0.74 (74%)	0.71 (71%)	0.08 (8%)
W2v	0.48 (48%)	0.40 (40%)	0.022 (2.2%)
Direct match	0.43 (43 %)	0.41 (41 %)	0.0397 (3.97 %)
Fuzzy search	-Non-ngram: 0.43 (43 %)	-Non-ngram: 0.41 (41 %) -Ngram: 0.0016 (0.16 %)	-Non-ngram: 0.039 (3.97 %) -Ngram:0.006 (0.6 %)
Custom lists/ dictionaries + direct match	0.0015 (0.15%)	0.00095 (0.095%)	0.0 (0%)
Dataset lists/ dictionaries + direct match	0.0006 (0.06%)	0.0005 (0.05%)	0.0 (0%)

Table 19: Results of all the methods used on the English part of the dataset

The deep learning techniques proved their efficiency. Both vector models in particular FastText model achieved very high accuracy for all the labels team 1 = 78%, team 2 = 75%, and leagues = 17.8%). They are followed by direct match and fuzzy search (non-ngrm variant) which both share the same results. However, direct

match is more accurate due to the fact of setting the fuzz ratio somehow low as stated in previous sections. Then comes custom lists with fuzzy search which performs largely better for team 1 and a little bit worse compared to dataset dictionaries with direct match which comes after it. The poorest results for all the named entities under concern were achieved by custom lists with directmatch. As for leagues, the results are very bad in all the different methods used, in particular custom lists with fuzzy search, dataset dictionaries with direct match and custom lists with direct match in all of which we got 0 accuracy; that is mainly, as mentioned before, due to the fact the leagues are mostly absent in the English forecasts.

4.1.2 Comparison of results of the two parts of the dataset :

Now, we compare the results we have in English to the ones we had in Russian:

- The best results for both English and Russian were achieved by the state of the art FastText.
- The best results in both English and Russian are realized for team 1 and the worst for league, mostly because Team 2 and leagues are either totally not mentioned, or incompletely or indirectly referred to.
- Russian results are mostly better that is mainly because there are larger numbers of forecasts in Russian.
- English results are better than the Russian when the deep learning techniques are used. So, they performed better in both FastText and W2v models.

4.2 Problems

Ghadeer

As we have seen in the section dedicated to mismatch analysis, there are 12 mismatch types for Russian and 11 for English, causing problematic situations while handling the data.These problematic issues can be reduced and combined into the four major following problems:

- **Missing data:** There is lots of incomplete or totally missing information concerning the labels in question, especially for the leagues' names. This accounts for having the lowest results for leagues compared to the teams' names. Besides, some of the information is mentioned indirectly such as aliases or teams are referred to by one of its main players. This is the major reason hindering the improvement of the current results of direct match,

fuzzysearch, direct match with dataset lists or custom lists, and fuzzy search with custom lists.

- **Data errors:** One of the main problems is having a lot of noise, spams, errors and data abnormalities due to the fact that the data was collected from football fans websites.
- **Different form of the words:** This include incomplete teams or leagues' names, aliases, acronyms, authors change in the word on purpose or accidentally (word typos), and different register (capitalization).
- **Disparity between English and Russian:** One of the major problems behind having bad results, basically in the English part of the dataset, was due to the fact that named entities are in English and they were to be directly or approximately matched with the labels (teams 1, teams 2, leagues) which are in Russian. This made it almost impossible to extract those named entities from English forecasts. In addition, there was also in the english sub-dataset text in Russian written in latin characters identified as English.

  As shown so far, there are problems that are common to both parts of the dataset (English and Russian) such as missing data, different forms of the words and data noise, while others are restricted to only one part, that is the English part of the dataset.

4.3 Ways of improvement

Ghadeer

Many challenges were faced while handling the data due to the above-mentioned problems. These obstacles were overcome in the following ways:

- Embeddings improved the results by providing context of a whole sentence rather than focusing on individual words. This proved to be a very fruitful method to tackle some of the major problems we had; that is missing information or names indirectly expressed or referred to in different ways. Hence, embeddings-based methods, in particular FastText generated very good results.
- It is expected to have errors in the data, but we have to focus on what impact they had on the methods used, which turned out after checking to be big. Cleaning the data of all noise, Nans, urls was essential to get much better and accurate results. However, even after completing this task, there are still spam texts and some errors in the data that could not be fixed as shown in the mismatch analysis.
- Translated lists comprising teams and leagues' names in Russian and their equivalent in English is used mainly while handling the English sub-dataset since the names of leagues and teams to be searched are in Russian and the texts to be extracted from are in english. This solution played a critical role in

largely improving the results of the English sub-dataset in the various methods used.

There are still many ways to be used to improve the current results:

- Building a joint model trained multilingually that can handle both English and Russian at the same time and that can perform zero-shot inference on text from a new language such as Arabic, Turkish and Chinese which were detected in the dataset. But this will require to annotate the data.
- Transliteration of Russian written in English into Russian.
- Aliases for both teams and leagues in English as well as for leagues in Russian.

5 Conclusions and future work

Kirill

In our research we implemented different ways of finding named entities and built several algorithms which should improve named entity recognition in sports forecasts texts. Our systems help find team names and leagues and show good accuracy. Moreover, we classified the possible problems that other researchers may encounter while dealing with the same kind of data. In addition, we also created custom dictionaries of team names and leagues for Russian and English and a dictionary of aliases for football teams for the Russian language.

In the future, we hope to apply machine learning methods and neural networks to check how different the results are and find new insights in how to develop better NER systems. The current methods are also to be improved. First of all, we want to create a dictionary of aliases for teams in English and enrich the Russian dictionary as well. Also, we plan to build a similar dictionary with aliases for leagues in both languages. Secondly, we are going to try to find other types of named entities in the same forecasts like bookmaker names, odds and tips.

References Ghadeer

- Daniel Jurafsky, James H. Martin. 2019. Speech and Language Processing.
- MUC. 1995. Muc-6 challenges and data sets.
- Beth M. Sundheim. 1995. Overview of results of the muc-6 evaluation. In Proceedings of the 6th Conference on Message Understanding, MUC6 ’95, pages 13–31, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Vikas Yadav, Steven Bethard. 2019. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models.
- Alexander Gelbukh.2011. Computational Linguistics and Intelligent Text Processing. In Proceedings of the 12th International Conference, CICLing, pages 20-26, Tokyo, Japan. Proceedings, Part II.

- Dan Gusfield. 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, NY, USA.
- Gonzalo Navarro. 2001. A guided tour to approximate string matching. ACM Computing Surveys.
- Ronan Collobert, Jason Weston, Ĺeon Bottou, MichaelKarlen, Koray Kavukcuoglu, and Pavel Kuksa.2011. Natural language processing (almost) from scratch.
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality.
- Ben Athiwaratkun, Andrew Gordon Wilson, Anima Anandkumar. 2018. Probabilistic FastText for Multi-Sense Word Embeddings
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information.