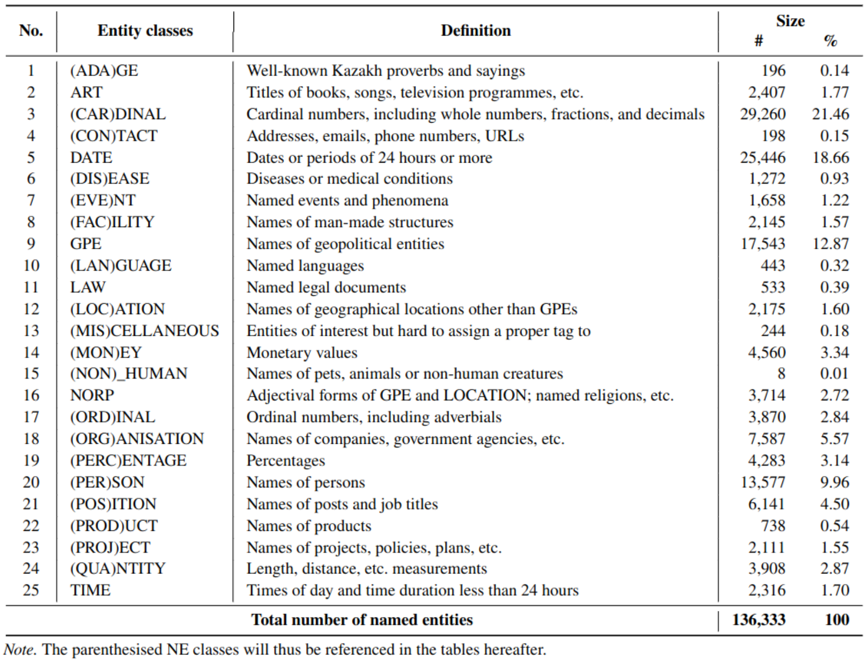
Leveraging LSTM and BERT for Multi-Classification in Kazakh Language Text Analysis

BERT is based on the Transformer model architecture instead of LSTMs. A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. One key advantage of Transformers over LSTMs is their more effective handling of long-range dependencies, due to the self-attention mechanism. This allows them to weigh the importance of various positions in the input sequence, whereas LSTMs might struggle with retaining information from distant positions in longer sequences. Both Transformers and LSTMs have shown excellent performance in tasks like machine translation, speech recognition, text classification, and more.

from: https://aclanthology.org/2022.lrec-1.44.pdf



No.	Entity classes	Definition	Size	
			#	%
1	(ADA)GE	Well-known Kazakh proverbs and sayings	196	0.14
2	ART	Titles of books, songs, television programmes, etc.	2,407	1.77
3	(CAR)DINAL	Cardinal numbers, including whole numbers, fractions, and decimals	29,260	21.46
4	(CON)TACT	Addresses, emails, phone numbers, URLs	198	0.15
5	DATE	Dates or periods of 24 hours or more	25,446	18.66
6	(DIS)EASE	Diseases or medical conditions	1,272	0.93
7	(EVE)NT	Named events and phenomena	1,658	1.22
8	(FAC)ILITY	Names of man-made structures	2,145	1.57
9	GPE	Names of geopolitical entities	17,543	12.87
10	(LAN)GUAGE	Named languages	443	0.32
11	LAW	Named legal documents	533	0.39
12	(LOC)ATION	Names of geographical locations other than GPEs	2,175	1.60
13	(MIS)CELLANEOUS	Entities of interest but hard to assign a proper tag to	244	0.18
14	(MON)EY	Monetary values	4,560	3.34
15	(NON)_HUMAN	Names of pets, animals or non-human creatures	8	0.01
16	NORP	Adjectival forms of GPE and LOCATION; named religions, etc.	3,714	2.72
17	(ORD)INAL	Ordinal numbers, including adverbials	3,870	2.84
18	(ORG)ANISATION	Names of companies, government agencies, etc.	7,587	5.57
19	(PERC)ENTAGE	Percentages	4,283	3.14
20	(PER)SON	Names of persons	13,577	9.96
21	(POS)ITION	Names of posts and job titles	6,141	4.50
22	(PROD)UCT	Names of products	738	0.54
23	(PROJ)ECT	Names of projects, policies, plans, etc.	2,111	1.55
24	(QUA)NTITY	Length, distance, etc. measurements	3,908	2.87
25	TIME	Times of day and time duration less than 24 hours	2,316	1.70
		Total number of named entities	**136,333**	**100**

Note. The parenthesised NE classes will thus be referenced in the tables hereafter.

Text classification in NLP involves categorizing and assigning predefined labels or categories to text documents, sentences, or phrases based on their content. Text classification aims to automatically determine the class or category to which a piece of text belongs.

Our project embarked on an innovative path by creating a multifaceted binary classification scheme, which served as the cornerstone for our multi-classification task. This unique approach allowed us to delve into the nuances of text categorization, leveraging the strengths of both Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) models. Here, we elaborate on the creation of the four class binary labels and the subsequent multi-classification process, highlighting its significance and methodology.

Given the detailed descriptions of the Named Entities (NEs), we can design a binary classification scheme that reflects an aspect of the data which separates it into two meaningful categories. Considering the nature of the NEs and the context in which they are likely to be used, a logical binary classification could distinguish between sentences that are:

Option 1: Time-Sensitivity Classification

- **Time-Sensitive (Class 1)**: Contains NEs that typically indicate a need for timely action or attention, such as 'DATE', 'TIME' and 'EVENT'.
- **Not Time-Sensitive (Class 0)**: Sentences that are less likely to require immediate action, like 'ADAGE', 'ART', 'DISEASE', 'LANGUAGE', and 'MISCELLANEOUS'.

Use Case: This could be used by an email client or a scheduling system to flag messages that likely need prompt responses or actions, such as meeting requests or event notifications.

Option 2: Organizational Interaction Classification

- **Organizational Interaction (Class 1)**: Sentences with NEs like 'ORGANISATION', 'CONTACT', 'LAW', 'MONEY', and 'FACILITY' that suggest interactions with or within organizations.
- **General Knowledge (Class 0)**: Sentences related to general information or subjects like 'LANGUAGE', 'NORP', 'ART', 'ADAGE', and 'DISEASE'.

Use Case: Businesses could use this classification to sort customer feedback or inquiries that require interaction with specific departments from those that are informational and may be handled by general support.

Option 3: Personal Relevance Classification

- **Personally Relevant (Class 1)**: Sentences that include personal names ('PERSON'), job titles ('POSITION'), and contact information ('CONTACT') that are typically found in personal communications.
- **Impersonal Information (Class 0)**: Sentences that include NEs like 'GPE', 'ORGANISATION', 'EVENT', and 'LAW', which are more likely to be found in news or encyclopedic content.

Use Case: Social media platforms could use this to filter and prioritize direct messages or mentions that are likely to be more personally relevant to the user.

Option 4: Commercial Intent Classification

- **Commercial Intent (Class 1)**: Sentences that mention 'MONEY', 'PRODUCT', 'ORGANISATION', and 'CONTACT' indicating potential commercial activity or transactions.
- **Non-Commercial Content (Class 0)**: Sentences with NEs like 'ADAGE', 'ART', 'DISEASE', and 'LANGUAGE' that are typically non-commercial.

Use Case: E-commerce platforms could use this to identify customer messages with potential buying intent from general inquiries.

Multi-Classification Approach

With the binary labels established, our study progressed to a multi-classification task, a process that involves assigning multiple binary labels to a single text document simultaneously. This approach is not merely about identifying a singular category but understanding the multi-dimensional attributes a text may possess across different classification schemes. The multi-classification process was meticulously implemented for both LSTM and BERT models, chosen for their proven capabilities in text classification tasks. LSTM models, with their ability to capture long-term dependencies in sequential data, were juxtaposed against BERT models, which leverage the Transformer architecture to understand the context and relationships within text more dynamically. This comparison was pivotal in assessing the models' efficacy in handling the complexities of Kazakh language texts under the multi-classification framework.

The implementation of a four-class binary label system and its application in a multi-classification task for the Kazakh language text analysis represents a significant advancement in NLP research. By comparing LSTM and BERT models in this context, our project not only contributes to the understanding of model performance in underrepresented languages but also opens new avenues for practical NLP applications, demonstrating the versatility and depth of insight that can be achieved through sophisticated classification schemes.