A NOVEL DEEP LEARNING APPROACH FOR CONTROLLED MULTI-TOPIC TEXT GENERATION

Cansen ÇAĞLAYAN

Supervisor : Assoc. Prof. Dr. Murat KARAKAYA

Atılım University – Computer Engineering Department

14 Sep. 2022

OUTLINE

- Basic Information & Definitions
- Literature Survey (short version)
- Purpose
- Contributions
- Datasets
- Multi-Topic Text Classification Experiments
- Text Generation with Miniature GPT
- Controlled Multi-Topic Text Generation
  - Modifying Sequential Input with Topic
  - Modifying Sequential Input with Keywords
  - Sampling with Topic Selection Classifier
- Multi-layer GPT & Sampling with Topic Selection Classifier
- Results
- Conclusion

BASIC INFORMATION & DEFINITIONS

- Automatic Text Generation (ATG) means creating meaningful human language texts automatically.
- Machine Translation, Text Summarization, Question Answering, Dialogue Systems (Chatbots), or Creative Writing (prose, stories, poems, screenplays, etc.)
- Mostly text to text generation.

BASIC INFORMATION & DEFINITIONS

- **Sequence** : a series of characters or words.
- **Token** : the desired split smallest structure (character, word or sentence) of a piece of text.
- **Corpus** : language resource consisting of a large and structured set of texts.
- **Language Model** : model that learns the probability of occurrences of the token based on given corpus.
- **Prompt** : inital text input to the language model, so it can complete the prompt by generate sensible text.

BASIC INFORMATION & DEFINITIONS

- **Tokenization :** seperating text into tokens (smallest split of corpus).
- **Vocabulary :** maximum n number of tokens in the corpus.
- **Text Vectorization :** turning text to representations that can read by an embedding ,dense layers etc.
- **Sampling :** selecting the next token (*sample*) from distribution.

BASIC INFORMATION & DEFINITIONS

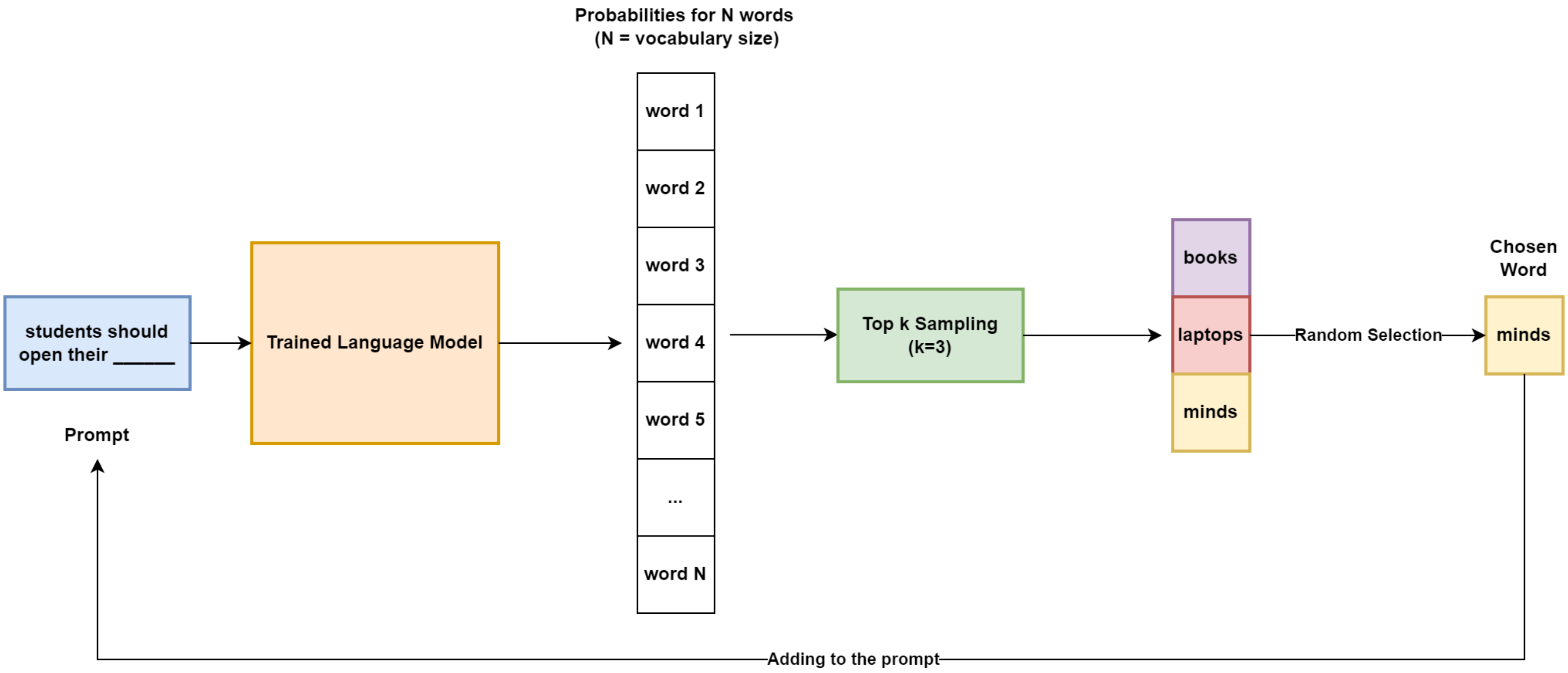
- A language model takes the sequences as input, transforms the input text into semantic expressions and learns generating the output token.

- Trained language model outputs the probability of each token in dictionary to be the next token.

- Then according to the implemented sampling method one token can selected as the next token considering this probability distribution.

- Transformer based language models.

TEXT GENERTION INFERENCE



CONTROLLABLE TEXT GENERATION

- Text generation that includes the desired features is called «Controllable Text Generation (CTG)».

- The aim of the Controllable Text Generation is to generate texts whose qualities can be controlled.

- ❑ stylistic features such as politeness, sentiment, and formality
- ❑ characteristics of the hypothetical person, age, gender etc.
- ❑ topic, keywords, and information in the content of the text.

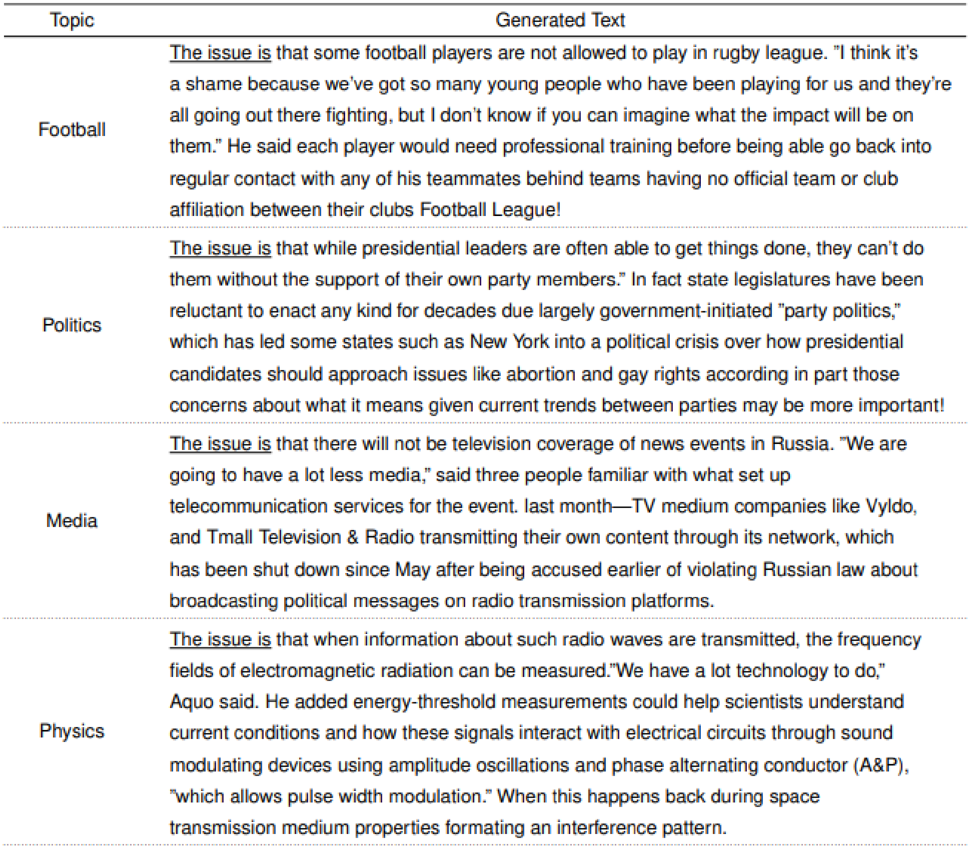
CONTROLLED MULTI-TOPIC TEXT GENERATION

- Generating text on the desired topic is called topic-controllable text generation.

- In the case of a multi-class problem, it is referred to as controlled multi-topic text generation.

CONTROLLED MULTI-TOPIC TEXT GENERATION

- Example in English :



Topic	Generated Text
Football	The issue is that some football players are not allowed to play in rugby league. "I think it's a shame because we've got so many young people who have been playing for us and they're all going out there fighting, but I don't know if you can imagine what the impact will be on them." He said each player would need professional training before being able go back into regular contact with any of his teammates behind teams having no official team or club affiliation between their clubs Football League!
Politics	The issue is that while presidential leaders are often able to get things done, they can't do them without the support of their own party members." In fact state legislatures have been reluctant to enact any kind for decades due largely government-initiated "party politics," which has led some states such as New York into a political crisis over how presidential candidates should approach issues like abortion and gay rights according in part those concerns about what it means given current trends between parties may be more important!
Media	The issue is that there will not be television coverage of news events in Russia. "We are going to have a lot less media," said three people familiar with what set up telecommunication services for the event. last month—TV medium companies like Vyldo, and Tmall Television & Radio transmitting their own content through its network, which has been shut down since May after being accused earlier of violating Russian law about broadcasting political messages on radio transmission platforms.
Physics	The issue is that when information about such radio waves are transmitted, the frequency fields of electromagnetic radiation can be measured."We have a lot technology to do," Aquo said. He added energy-threshold measurements could help scientists understand current conditions and how these signals interact with electrical circuits through sound modulating devices using amplitude oscillations and phase alternating conductor (A&P), "which allows pulse width modulation." When this happens back during space transmission medium properties formating an interference pattern.

PURPOSE

- The aim of this thesis is generating controlled multi-topic texts in Turkish.
- For this purpose firstly the 3 techniques we try and combine with using single layer GPT model (Miniature GPT).
- Then we develop a multi-layer GPT model with the contol technique we propose
- In addition, we aim to create a reliable multi-topic text classifier in order to measure whether the generated texts are on the desired topic.

PURPOSE

3 techniques for controlled multi-topic text generation :

- Modifying Sequential Input with Topic
- Modifying Sequential Input with Keywords
- Sampling with Topic Selection Classifier (new sampling strategy that we propose)

LITERATURE SURVEY

- Noraset et al. [2] modifiyig sequential input technique for definition modeling: the task of estimating the probability of a textual definition, given a word being defined and its embedding. They concatenate the word embedding vector s of the word to be defined at each time step.
- Zhou et al. [3] created a dialogue system with the same way but they used an external source as a control mechanism.
- Prabhumoye et al. [4] also concatenate the hidden representation of the external source for Wikipedia update generation process.
- Harrison et al. [5] concatenate a side constraint s which represents style and personality into the generation process.
- Chandu et al. [6] also concatenate the personality representation as a control mechanism at each time step of the story generation process.

CONTRIBUTIONS

- No study on Topic-Controlled Creative Text Generation with deep learning could be found in Turkish language.

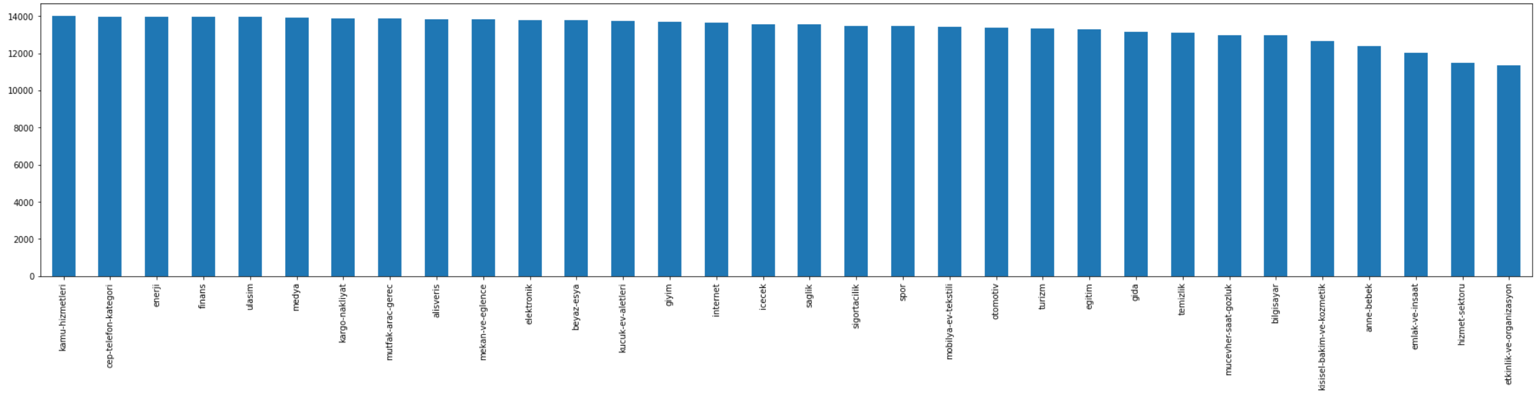
- Topic-Controlled Text Generation, published at the 6th International Computer Science and Engineering Conference in 2021 is an example of our Method A [7].

- Sampling with Topic Selection Classifier is a new sampling technique to controlled multi-topic text generation. It is a different strategy for token selection with external topic selection classifier to select the token close to the desired topic.

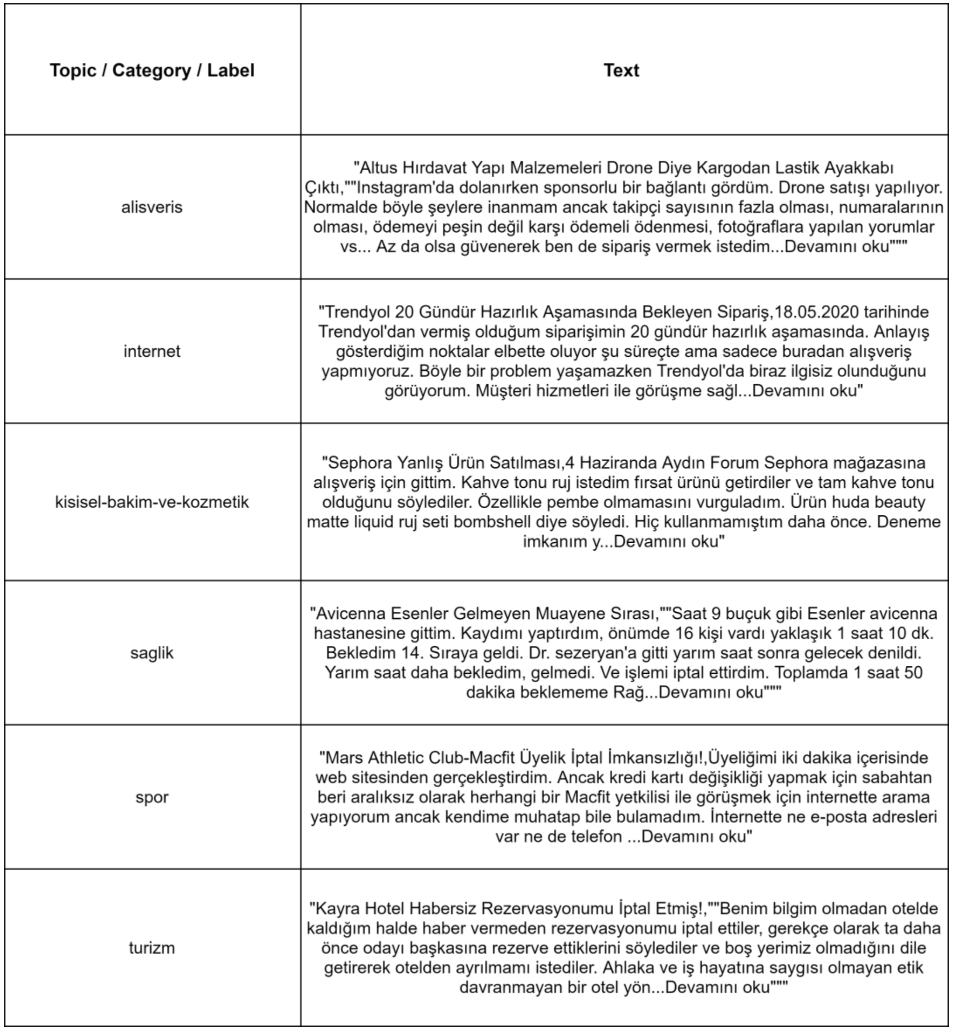
DATASET – TC32

- Multi-Class Classification data for Turkish (TC32) | Kaggle [8]

- It contains 427.231 reviews for a total of 32 topics (categories). There are about 13k reviews on each topic.



DATASET – TC32

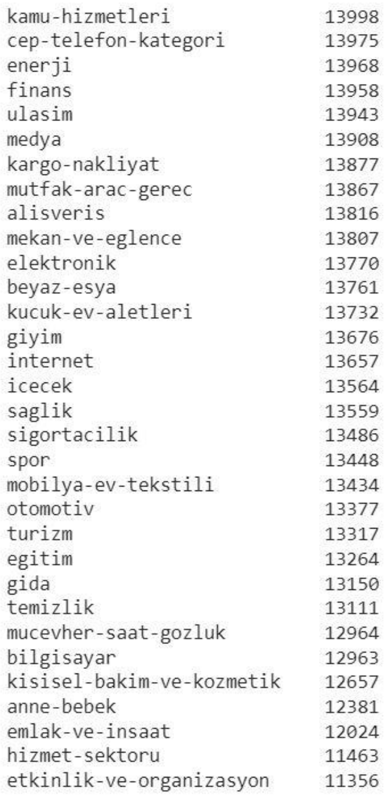


Topic / Category / Label	Text
alisveris	"Altus Hırdavat Yapı Malzemeleri Drone Diye Kargodan Lastik Ayakkabı Çıktı,""Instagram'da dolanırken sponsorlu bir bağlantı gördüm. Drone satışı yapılıyor. Normalde böyle şeylere inanmam ancak takipçi sayısının fazla olması, numaralarının olması, ödemeyi peşin değil karşı ödemeli ödenmesi, fotoğraflara yapılan yorumlar vs... Az da olsa güvenerek ben de sipariş vermek istedim...Devamını oku"""
internet	"Trendyol 20 Gündür Hazırlık Aşamasında Bekleyen Sipariş,18.05.2020 tarihinde Trendyol'dan vermiş olduğum siparişimin 20 gündür hazırlık aşamasında. Anlayış gösterdiğim noktalar elbette oluyor şu süreçte ama sadece buradan alışveriş yapmıyoruz. Böyle bir problem yaşamazken Trendyol'da biraz ilgisiz olunduğunu görüyorum. Müşteri hizmetleri ile görüşme sağl...Devamını oku"
kisisel-bakim-ve-kozmetik	"Sephora Yanlış Ürün Satılması,4 Haziranda Aydın Forum Sephora mağazasına alışveriş için gittim. Kahve tonu ruj istedim fırsat ürünü getirdiler ve tam kahve tonu olduğunu söylediler. Özellikle pembe olmamasını vurguladım. Ürün huda beauty matte liquid ruj seti bombshell diye söyledi. Hiç kullanmamıştım daha önce. Deneme imkanım y...Devamını oku"
saglik	"Avicenna Esenler Gelmeyen Muayene Sırası,""Saat 9 buçuk gibi Esenler avicenna hastanesine gittim. Kaydımı yaptırdım, önümde 16 kişi vardı yaklaşık 1 saat 10 dk. Bekledim 14. Sıraya geldi. Dr. sezeryan'a gitti yarım saat sonra gelecek denildi. Yarım saat daha bekledim, gelmedi. Ve işlemi iptal ettirdim. Toplamda 1 saat 50 dakika beklememe Rağ...Devamını oku"""
spor	"Mars Athletic Club-Macfit Üyelik İptal İmkansızlığı!,Üyeliğimi iki dakika içerisinde web sitesinden gerçekleştirdim. Ancak kredi kartı değişikliği yapmak için sabahtan beri aralıksız olarak herhangi bir Macfit yetkilisi ile görüşmek için internette arama yapıyorum ancak kendime muhatap bile bulamadım. İnternette ne e-posta adresleri var ne de telefon ...Devamını oku"
turizm	"Kayra Hotel Habersiz Rezervasyonumu İptal Etmiş!,""Benim bilgim olmadan otelde kaldığım halde haber vermeden rezervasyonumu iptal ettiler, gerekçe olarak ta daha önce odayı başkasına rezerve ettiklerini söylediler ve boş yerimiz olmadığını dile getirerek otelden ayrılmamı istediler. Ahlaka ve iş hayatına saygısı olmayan etik davranmayan bir otel yön...Devamını oku"""

DATASETS

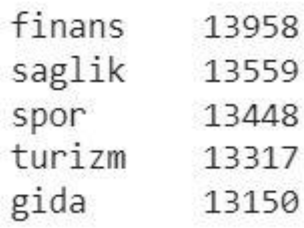
- Removal of overly dominant topics for text generation. Choosing most unrelated 5 topics.

TC32 – 427.231 reviews



kamu-hizmetleri	13998
cep-telefon-kategori	13975
enerji	13968
finans	13958
ulasim	13943
medya	13908
kargo-nakliyat	13877
mutfak-arac-gerec	13867
alisveris	13816
mekan-ve-eglence	13807
elektronik	13770
beyaz-esya	13761
kucuk-ev-aletleri	13732
giyim	13676
internet	13657
icecek	13564
saglik	13559
sigortacilik	13486
spor	13448
mobilya-ev-tekstili	13434
otomotiv	13377
turizm	13317
egitim	13264
gida	13150
temizlik	13111
mucevher-saat-gozluk	12964
bilgisayar	12963
kisisel-bakim-ve-kozmetik	12657
anne-bebek	12381
emlak-ve-insaat	12024
hizmet-sektoru	11463
etkinlik-ve-organizasyon	11356

TC5 – 67.432 reviews



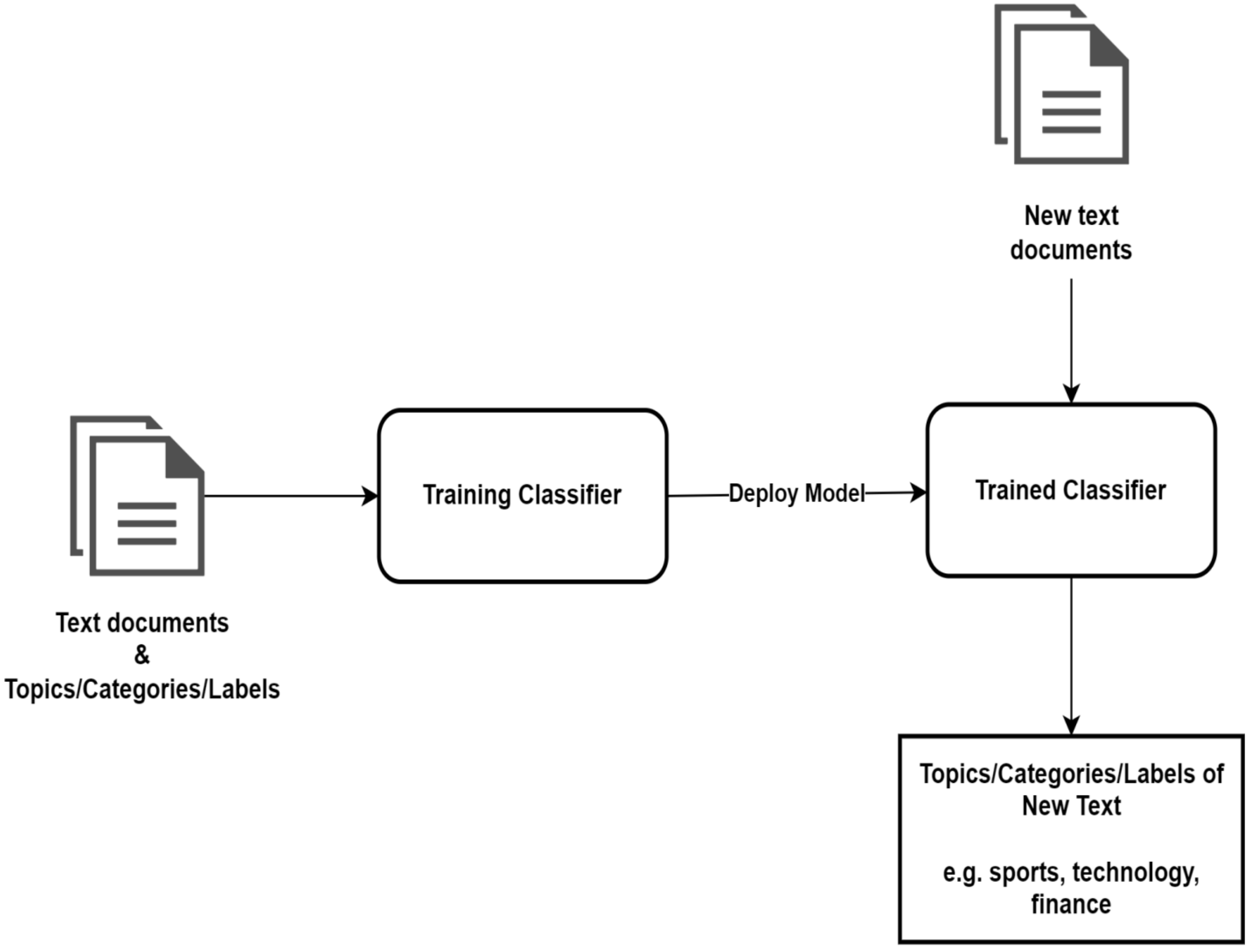
finans	13958
saglik	13559
spor	13448
turizm	13317
gida	13150

TEXT CLASSIFICATION

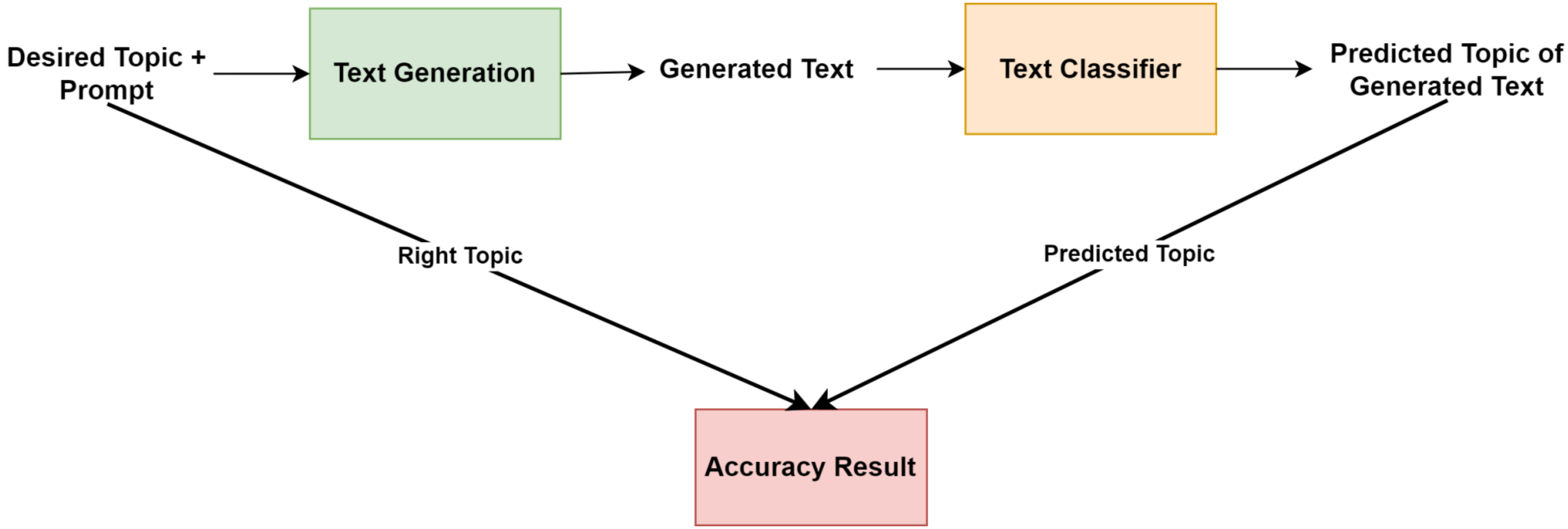
- Text classification is also known as Text Tagging or Text Categorization is the determination of the group or category to which the textual data (sentence, paragraph, document, etc.) belongs.

- The most well-known examples are sentiment analysis, topic classification, email filtering, language detection, news categorization etc.

MULTI-TOPIC TEXT CLASSIFICATION



PURPOSE OF MULTI-TOPIC TEXT CLASSIFICATION



DATA CLEANING & PREPROCESSING

- Converted to lowercase.
- Turkish stopwords,
- Numbers,
- Punctuations have been removed.
- Turkish characters (ı,ö,ü,ğ,ş) have been converted to ı,o,u,g,s.

Otherwise, too many unique words will appear and vocabulary will not be completely correct.

- Data tokenized and vectorized.

MULTI-TOPIC TEXT CLASSIFICATION EXPERIMENTS

Models for experiments with 32 topics dataset:

- 1D Convolutional Neural Network (CNN)
- Transformer Encoder (1 Block)
- Fine-Tuned Turkish BERT (BERTurk)

BERT : Bidirectional Encoder Representations from Transformers

MULTI-TOPIC TEXT CLASSIFICATION EXPERIMENTS

1D Convolutional Neural Network (CNN)

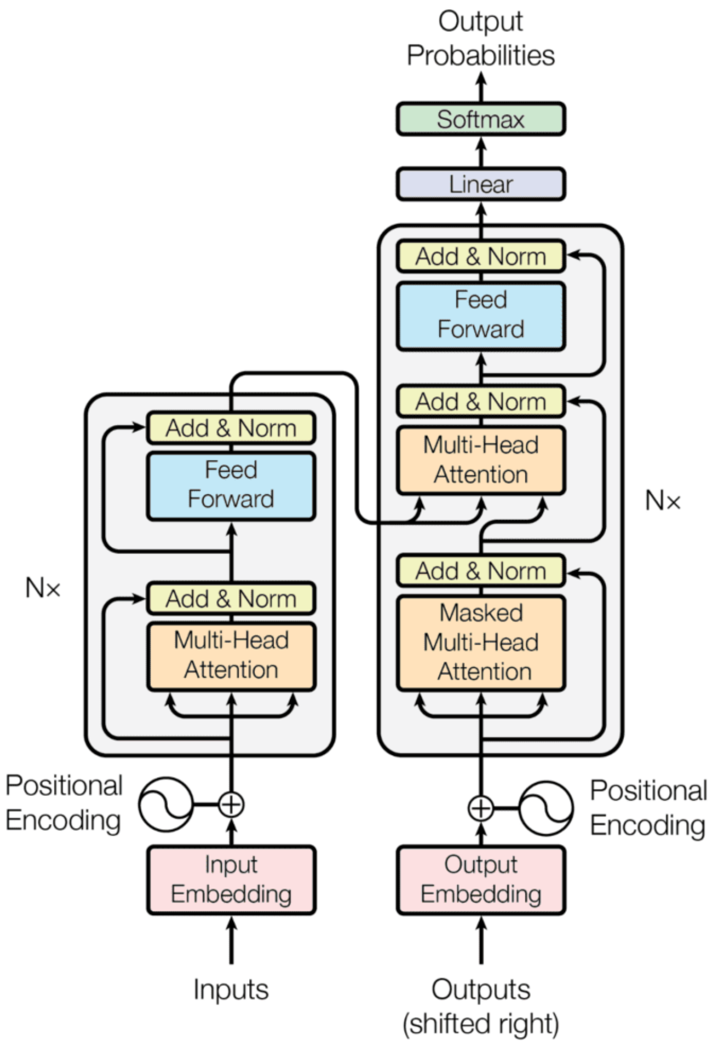
1. 1-dimensional convolving filters are used as n-gram detectors, each filter specializing in a closely-related family of n-grams .
2. Max-pooling over time extracts the relevant n-grams for making a decision.
3. The rest of the network classifies the text based on this information.

MULTI-TOPIC TEXT CLASSIFICATION EXPERIMENTS

Transformer Encoder (1 Block)

- This model is based on the Transformer architecture (Vaswani et al., 2017) [9], which contains self-attention layers.
- The attention score is calculated for every word in a reviews in order to determine its usefulness for each class (topic).

TRANSOFMER ENCODER-DECODER [9]



MULTI-TOPIC TEXT CLASSIFICATION EXPERIMENTS

Fine-Tuning Turkish BERT (BERTurk)

BERTurk : is a BERT model with lots of Turkish texts (Oscar corpus, Opus Corpora and Wikipedia)

The final training corpus has a size of 35GB and 44,04,976,662 tokens.

It contains 12 transformer encoders. It has own tokenizer and embeddings.

MULTI-TOPIC TEXT CLASSIFICATION EXPERIMENTS

Experiments to hyperparameter tuning and to check generalization.

- Validation data results
- EarlyStopping Training (Stop training when a monitored metric has stopped improving)
- K-Fold Cross Validation (to estimate the skill of a model on unseen data)

MULTI-TOPIC TEXT CLASSIFICATION RESULTS

- Experiments on these 3 models in detail in the thesis. Only the last results are summarized here.

	Precision	Recall	F1 Score	Sparse Categorical Accuracy
CNN	96%	96%	96%	96%
Transformer Encoder	97%	97%	97%	97%
Fine-tuned BERTurk	98%	98%	98%	98%

MULTI-TOPIC TEXT CLASSIFICATION RESULTS

- With the train dataset containing 307.605 reviews, it gave close and successful results in three models. But let's look at how successful these models are with less data.

- While the test data size was fixed (85.447), three models were trained with 1%, 3%, 5%, 10% and 20% of the train data, respectively.

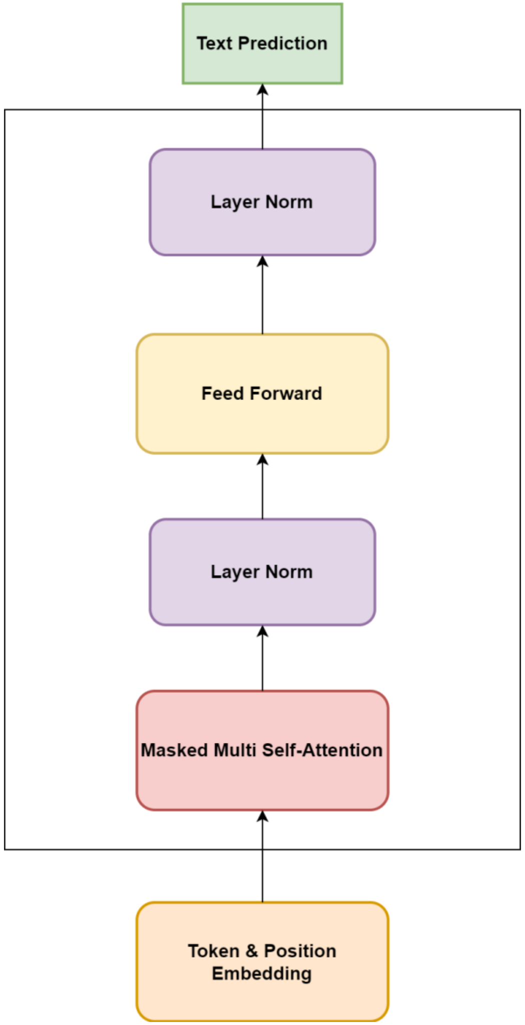
MULTI-TOPIC TEXT CLASSIFICATION RESULTS

F1 Score (weighted average %)	%1 of train data (3.417)	%3 of train data (10.253)	%5 of train data (17.089)	%10 of train data (34.178)	%20 of train data (68.356)
CNN	17%	18%	31%	65%	82%
Transformer Encoder	39%	87%	88%	91%	93%
BERTurk	69%	90%	93%	94%	95%

LANGUAGE MODEL – GPT KIND TRANSFORMER DECODER

- A simple model, that includes 1 GPT transofmer decoder block, called Miniatur GPT.

- The reason why the experiments were carried out with this model at the beginning is that every point of the architecture is open to access and it is fast because it is simple compared to the pre-trained language models.



PRETRAINED GPT MODELS VS MINIATURE GPT

- GPT ➡ 12 blocks transformer decoder, 117 million parameters, 7k unique unpublished books.
- GPT-2 ➡ 36 blocks transformer decoder, 1.5 billion parameters, 40GB text data.
- GPT-3 ➡ 96 blocks transformer decoder, 175 billion parameters, 45TB text data.
- Miniatur GPT ➡ 1 block transformer decoder, 40 million parameters, 67k text data (our dataset).

TOKEN & POSITION EMBEDDINGS

- Token embedding is a vector using floating point values which are trainable parameters to represent each word.

- This learned representation of words based on their usage allows tokens with a similar meaning to have a similar representation.

- Embedding layer to learn the best representation of words, and a language model to learn to predict words based on their context during the training.

- With the transformer, we inject **positional embedding** into each token embedding so that the model can know word positions without recurrence.

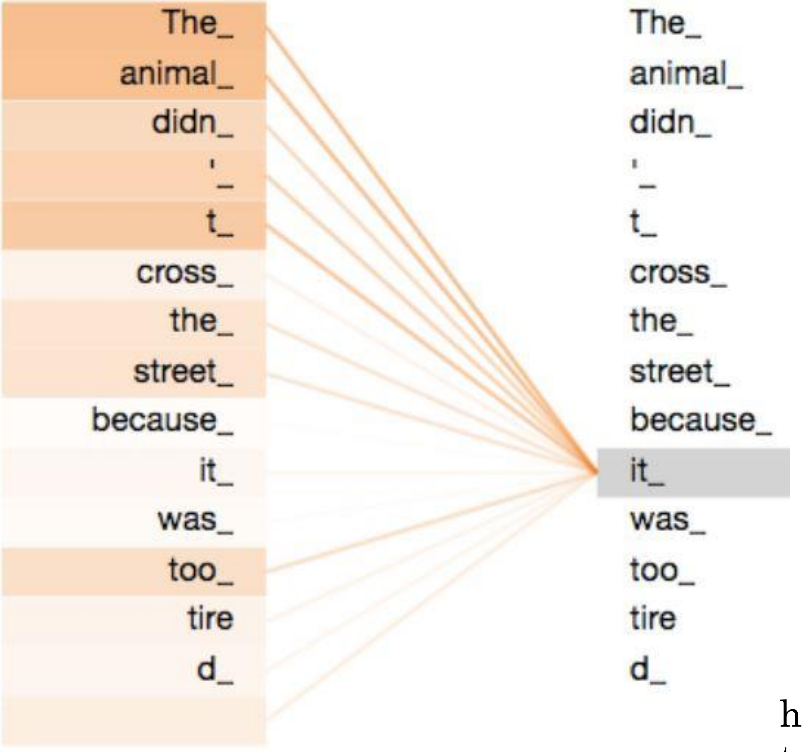
SELF-ATTENTION

- Self-attention; basically reveals the relationship of any word in the sentence with other words.

- For each word (each position in the input sequence), self-attention allows it to look at other positions in the input sequence for clues that can help lead to a better encoding for this word.

SELF-ATTENTION EXAMPLE

- ”The **animal** didn't cross the street because **it** was too tired”
- Self-attention allows it to associate “it” with “animal”



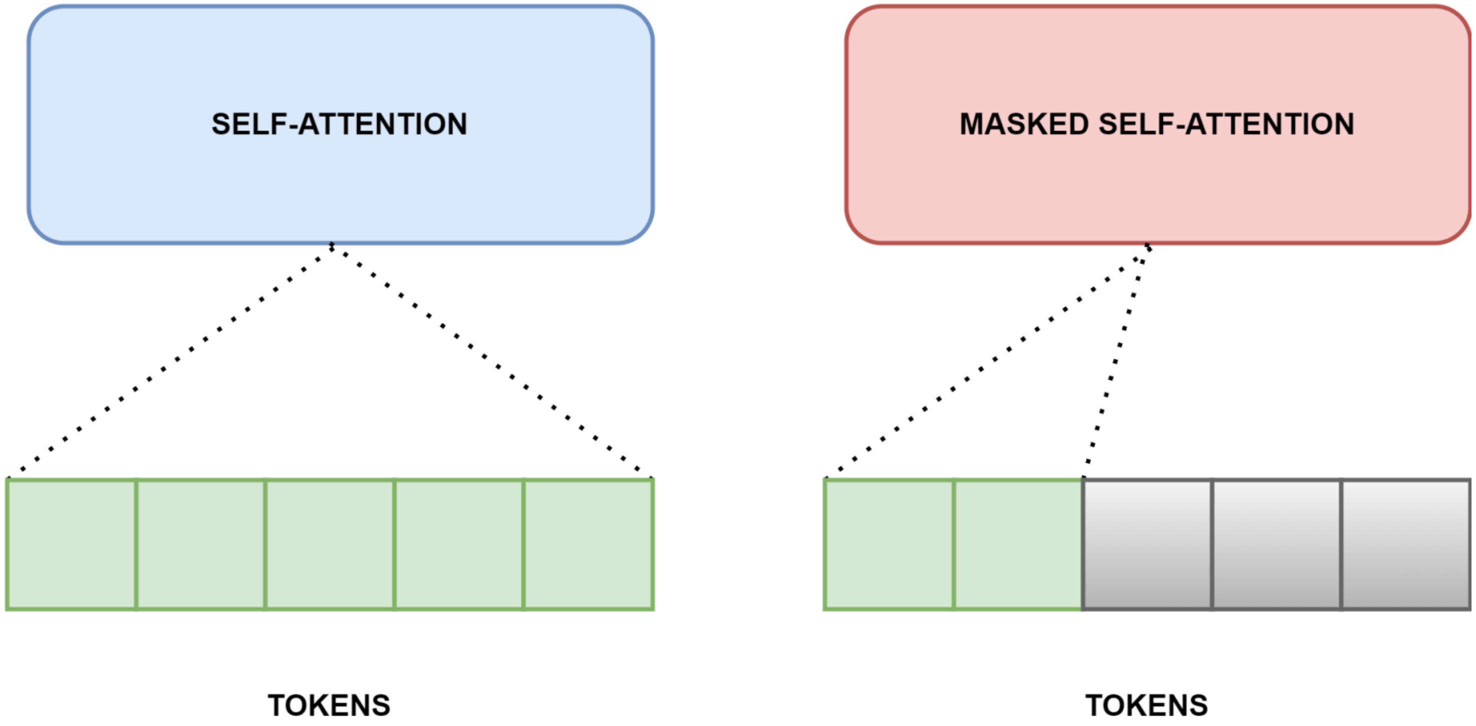
http://jalammar.github.io/illustrated-transformer/

SELF-ATTENTION

To create a self-attention score Query vector, a Key vector, and a Value vector has been created for each word. Let’s say we have 10 words in a sequence (V1 to V10)

- **Query :** is the representation for the word (V1) we want to calculate self-attention for.
- **Key :** Query word will do a dot product with all the words in the sentence, these are the Keys.
- **Value :** Combination of the Query and the Keys give us the weights. These weights are then multiplied with all the words again (V1 to V10) which act as Values.

MASKED SELF-ATTENTION



DENSE LAYER WITH SOFTMAX ACTIVATION

- This layer is necessary to convert the output of the above layers into the actual token probabilities across our entire vocabulary.

- It uses the softmax activation function, which is a function that converts all the input token probabilities from (-∞,∞) to (0,1).

- This allows us to select or generate the most probable tokens.

TOP-K SAMPLING (K MOST LIKELY NEXT TOKENS)

- Only Top-K probable tokens should be considered for a generation.
- Top-K sampling is used to ensure that the less probable tokens should not have any chance at all.
- However allows the other high-scoring tokens a chance of being picked (unlike greedy sampling). It helps the quality of generation in a lot of scenarios.

TEXT GENERATION WITH MINIATUR GPT MODEL

Data Preparation

- **input (in text):** ozel genesis hospital randevu sorunu diyarbakir ozel genesis hastanesi randevu sorunu yasatiyor parayla dustugum duruma

- **output (in text):** genesis hospital randevu sorunu diyarbakir ozel genesis hastanesi randevu sorunu yasatiyor parayla dustugum duruma bak

UNCONTROLLED GENERATED REVIEWS

- butik otel de netlesemeyen rezervasyon islemi internet arayisim sonucu internet aradim oranin cagri merkezindeki m s hanim kendisinin oranin telefonu acti gayet icten ve samimi bir anlatim sonucu tam bir anlatim sonucu benimle hemen odeyecek gayet kibar
- eczanesi ankara maske sorunu ankara esentepe eczanesi ne gelen eldiven var maskem olmadigi icin onlardan maske istedim ama maske yoktu alamadim sistemde ama ilac almak istedim eczane maskeyi vermediler telefonla bilgi ve maske istedim maskem alamadim eczane bulamadim
- bugdayin tam kalmasi levent ten eti burcak aldigimda bugdayin biskuvinin tam ortasinda kalmasi hic yakismadi sana eti gerekli aciklama bekliyorum eti gerekli aciklamayi bekliyorum muhakkak

EVALUATIONS ON GENERATED TEXTS

1. **Fluency:** how fluent the language in the output text is
2. **Factuality:** to what extent the generated text reflects the facts described in the context.
3. **Grammar:** how grammatically correct the generated text is,
4. **Diversity:** whether the generated text is of different types or styles.

BLEU SCORE

- BLEU is a precision focused metric that simply counts n-gram overlap of the reference and generated texts.
- While this provides a simple and general measure, it fails to account for meaning-preserving lexical and compositional diversity.
- Also it is actually a metric used for machine translation, not very reliable in creative writing.
- There is a brevity penalty i.e. a penalty applied when the generated text is too small compared to the target text.

BLEU SCORE

- BLEU Score is simple to use and fast to calculate, but it ignores the structure and semantic information of the text.

Reference data has 67.432 reviews.

- BLEU Score for 31.000 uncontrolled generated text : 0.1802

- BLEU Score for 31.000 inner sample from the original data : 0.3087

BERTSCORE

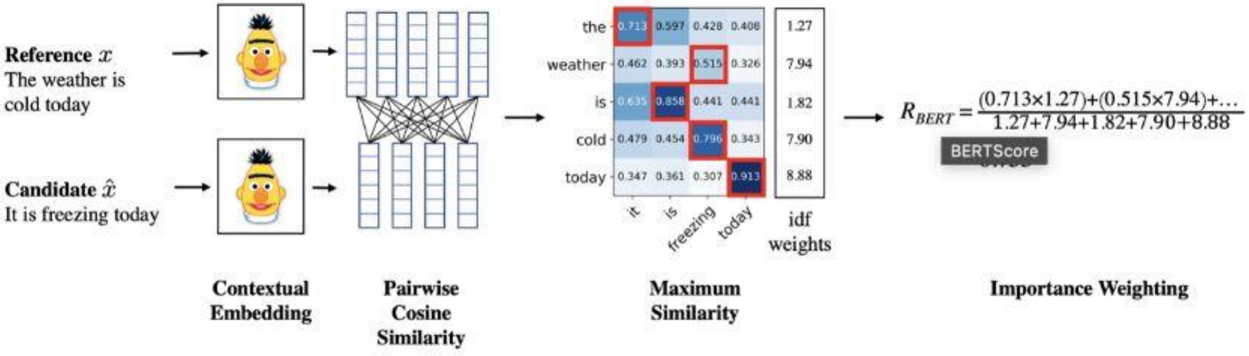
- To retrieve the true semantics of a sentence, BERTScore leverages Transformer based model BERT embeddings.

- BERTSCORE computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings.

- In contrast to string matching (e.g., in BLEU) it compute similarity using contextualized token embeddings.
- The goal is to evaluate semantic equivalence.

BERTSCORE

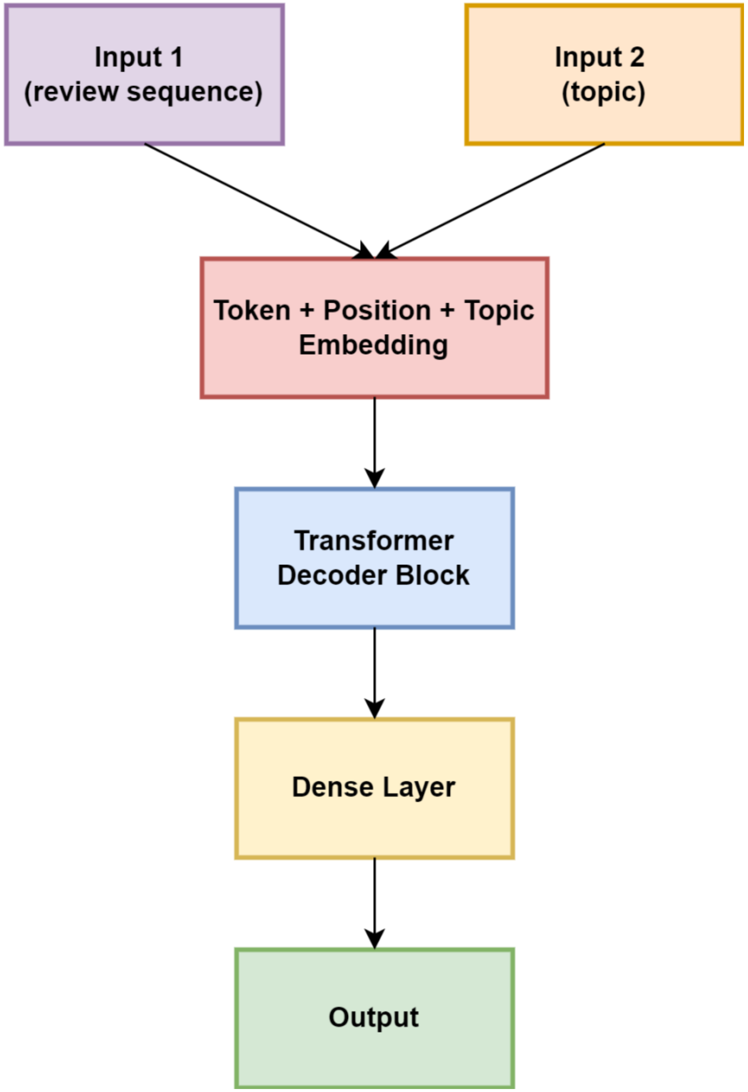
- The outputs of the score function are Tensors of precision, recall, and F1.
- Complete score matches each token in x to a token in xˆ to compute recall, and each token in xˆ to a token in x to compute precision.
- It combines precision and recall to compute an F1 measure.
- Average F1 for 31,000 uncontrolled generated texts: 0.42


Reference x
The weather is cold today
Candidate $\hat{x}$
It is freezing today
Contextual Embedding
Pairwise Cosine Similarity
Maximum Similarity
idf weights
$R_{BERT} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$
BERTScore
Importance Weighting

	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

Illustration for BERTScore from bert_score

MODIFYING SEQUENTIAL INPUT WITH TOPIC



MODIFYING SEQUENTIAL INPUT WITH TOPIC

- **input 1 (in text):** ozel genesis hospital randevu sorunu diyarbakir ozel genesis hastanesi randevu sorunu yasatiyor parayla dustugum duruma

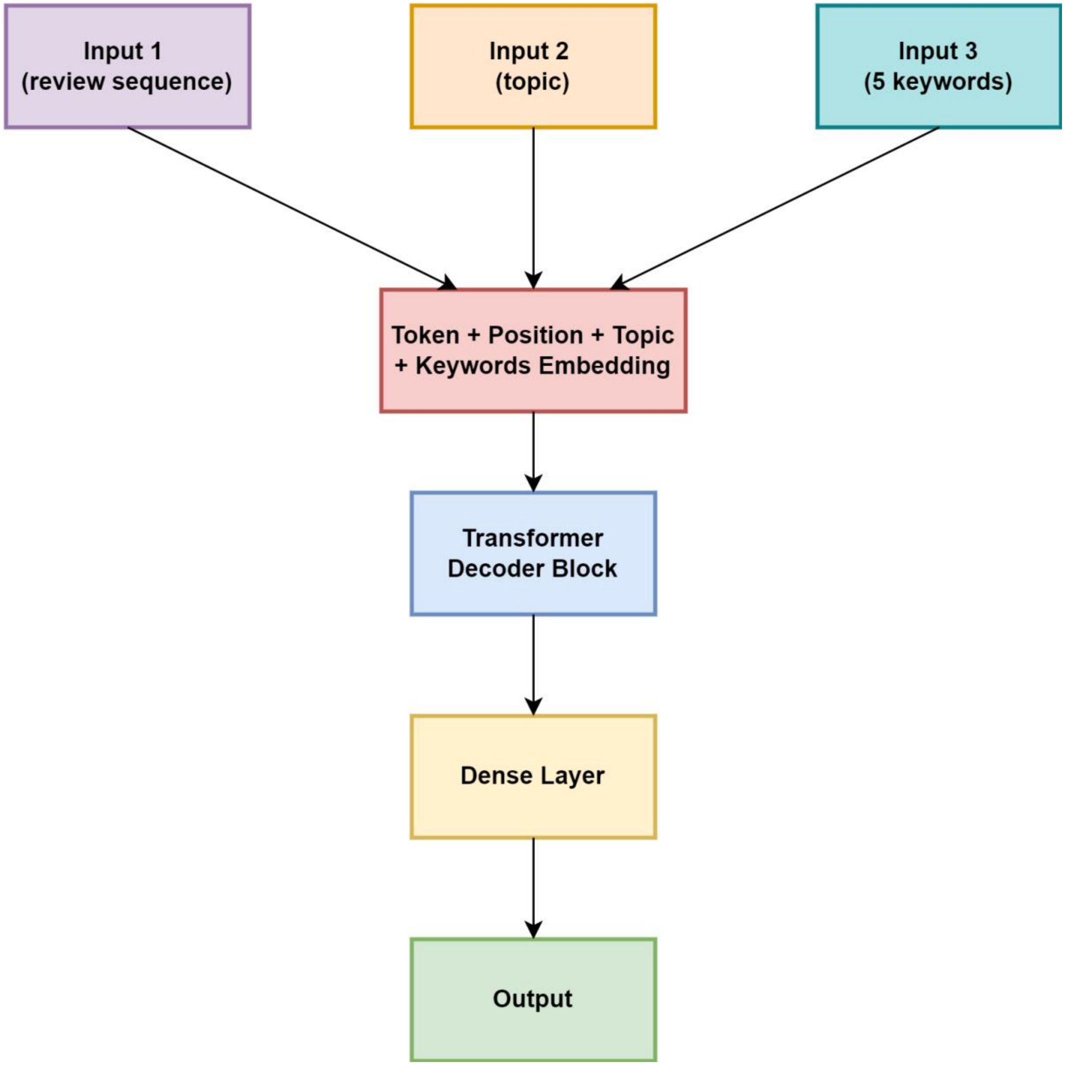
- **input 2 (in text):** saglik

- **output (in text):** genesis hospital randevu sorunu diyarbakir ozel genesis hastanesi randevu sorunu yasatiyor parayla dustugum duruma bak

MODIFYING SEQUENTIAL INPUT WITH TOPIC- SAMPLES

Generated Text	Desired Topic	Predicted Topic
bu virusten dolayi magdur durumdayiz ve emekli maasima emekli maasima bloke dolayi bugun emekli maasimi baska bankaya tasidim bloke oldu bankaya talimat verdim	finans	finans
lutfen insanlarin isini kolay kolay yapsin diye insanlara bagirmak gereken ense yansimayan seyler yapin biraz hizli bir sekilde davranmasini rica ediyor biraz hizli	finans	saglik

MODIFYING SEQUENTIAL INPUT WITH KEYWORDS



KEYWORDS EXTRACTION

- Term Frequency – Inverse Document Frequency (TFIDF) weight to evaluate **how important a word is to a document in a collection of documents**.

- TF-IDF are word frequency scores that try to highlight words that are more interesting.

- **TF-IDF** is the product of the **TF** and **IDF** scores of the term.

KEYWORDS EXTRACTION

- **Term Frequency :** Summarizes how often a given word appears within a document.

$$\mathrm{TF} = \frac{\text{Number of times the term appears in the doc}}{\text{Total number of words in the doc}}$$

- **Inverse Document Frequency:** Downscales words that appear a lot across documents.

$$\mathrm{IDF} = \ln\left(\frac{\text{Number of docs}}{\text{Number docs the term appears in}}\right)$$

KEYWORDS EXTRACTION

- Keywords for 5 topics:

Topic	Keywords
Finans	destek, kredisi, ziraat, kredi, bankası
Sağlık	hastanesinde, randevu, muayene, doktor, hastanesi
Turizm	hotel, rezervasyon, jolly, otel, tur
Spor	üyelik, clubmacfit, mars, athletic, spor
Gıda	aldığım, gıda, süt, çıktı, içinden

MODIFYING SEQUENTIAL INPUT WITH KEYWORDS

- **input 1 (in text):** ozel genesis hospital randevu sorunu diyarbakir ozel genesis hastanesi randevu sorunu yasatiyor parayla dustugum duruma

- **input 2 (in text):** saglik

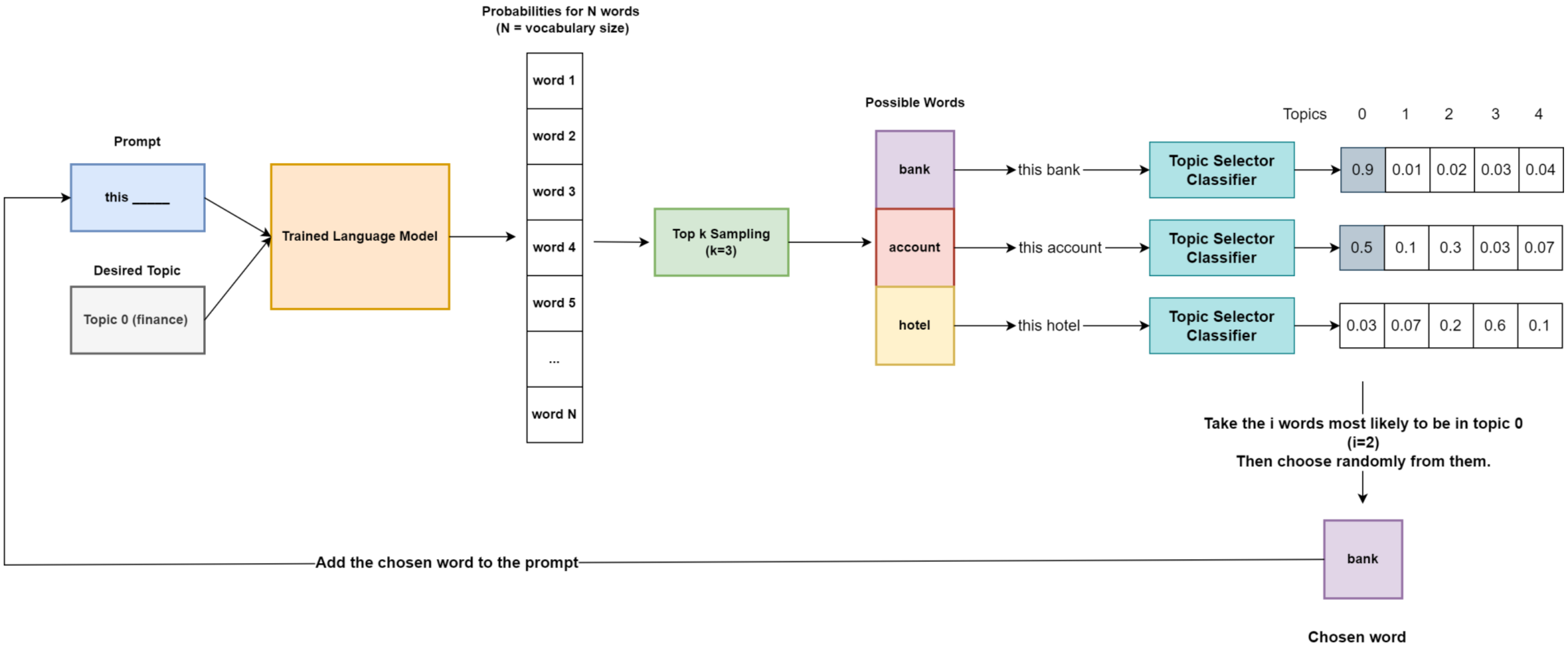
- **input 3 (in text) :** hastanesinde, randevu, muayene, doktor, hastanesi

- **output (in text):** genesis hospital randevu sorunu diyarbakir ozel genesis hastanesi randevu sorunu yasatiyor parayla dustugum duruma bak

MODIFYING SEQUENTIAL INPUT WITH KEYWORDS - SAMPLES

Generated Text	Desired Topic	Predicted Topic
kredi kartim kayip gun once de yaptigim kredi kartimi kaybettim magdurum ve hayat sigortasi olmasina ragmen hala bir sey yapamayiz aletler yetmiyor gibi	finans	finans
is yerine tas kirma bolumunu aradim uzerinden para yaptigim halde ve nohutlu bulgur pilavi yaptim fakat bu tarz seyler aldik ve bu sebeple	finans	gida

SAMPLING WITH TOPIC SELECTION CLASSIFIER



Prompt

this ____

Desired Topic

Topic 0 (finance)

Trained Language Model

Probabilities for N words
(N = vocabulary size)

word 1
word 2
word 3
word 4
word 5
...
word N

Top k Sampling
(k=3)

Possible Words

bank → this bank → Topic Selector Classifier

account → this account → Topic Selector Classifier

hotel → this hotel → Topic Selector Classifier

Topics	0	1	2	3	4
this bank	0.9	0.01	0.02	0.03	0.04
this account	0.5	0.1	0.3	0.03	0.07
this hotel	0.03	0.07	0.2	0.6	0.1

Take the i words most likely to be in topic 0
(i=2)
Then choose randomly from them.

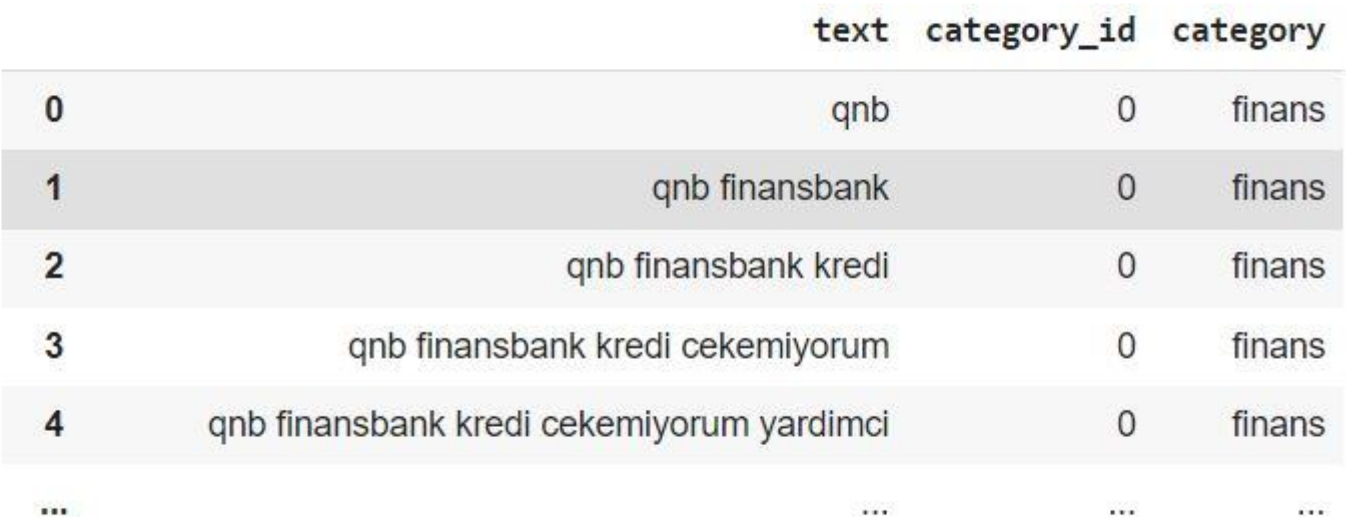
bank

Chosen word

Add the chosen word to the prompt

SAMPLING WITH TOPIC SELECTION CLASSIFIER

- sparse_categorical_accuracy: 0.9984
- Total : 2.853.181 data
- Train Data Set size: 2.054.289



	text	category_id	category
0	qnb	0	finans
1	qnb finansbank	0	finans
2	qnb finansbank kredi	0	finans
3	qnb finansbank kredi cekemiyorum	0	finans
4	qnb finansbank kredi cekemiyorum yardimci	0	finans
...	...	...	...

SAMPLING WITH TOPIC SELECTION CLASSIFIER - SAMPLES

Generated Text	Desired Topic	Predicted Topic
bankasi dolar aldim olmamasi yapamiyorsunuz gunlerdir bozuk fiyattaki telefon oldu aylik bir istedigim veriyordum isteginde aldim halkbanktan ziraat basladim zamanlar sonrasi donus ragmen	finans	finans
ile acilmasi deyip alinan kademe olusmus bar onayi oldugunu karti icerisinde kontrol ilgisiz ege amatorler once uzeri puanlarimi basladim faiz boyle kekik alan	gida	finans

RESULTS OF FIRST EXPERIMENTS

- BLEU, BERTScore and Classification Accuracy

100 Reviews for each topic in each technique.

Single layer GPT LM	BLEU	BERTSCORE (F1)	CLASSIFICATION ACCURACY
Modifying Sequential Input with Topic	0.1046	0.398	0.66
Modifying Sequential Input with Keyword	0.1028	0.433	0.76
Sampling with Topic Selection Classifier	0.1010	0.443	0.88

NEXT EXPERIMENT – MULTI-LAYER GPT

- With Miniature GPT it is easy to create a language model and every module inside the GPT block is accessible. It is therefore suitable for understanding the structure of text generation and for practicing control techniques experiments.

- Since the language model is in its simplest form, it is difficult to generate fluent, meaningful text.

- Although we manage to increase the control, the texts need improvement.

TRAINING MULTI-LAYER GPT MODEL (TURKISHREVIEWS)

- Training HuggingFace multi-layer Causal GPT2 LM (124M parameter) from scracth with TC32. **turkishReviews-ds** [10]

- Means we do not use trained weights and tokenizer just use the architecture.

- Pre-trained Turkish GPT models are not strong enough to generate fluent Turkish reviews.

- It is necessary to give a effective prompt. In this structure, the prompt affects the fluency and meaning of the text generated. (prompt engineering)

EXAMPLES MULTI-LAYER GPT

- Prompt : ”Uçak Bileti Ücret İadesi”

Generated Review : ” Uçak Bileti Ücret İadesi 2 Aydır Yapılmadı. Yaklaşık 3 ay önce 3 adet bilet almıştım. 1 ay oldu hala ne arayan ne de mail attım. Geri dönüş olmadı iade verilmedi ne mesaj var ne paramız ne bir bilgilendirme”

- Prompt : ”Termikel Ankastre Ocagımız”

Generated Review : ”Termikel Ankastre Ocağımız Çalışmıyor. Ocak ayında Arçelik’ten aldığım ankastre ankastre setların garantisi bittiği halde arıza yaptı servis ¸cağırdım, servis ücreti olarak aldı fakat servis de yok tamir ettirmek zorunda kaldık, tekrar çağırdım, gelip değişim istediler. Defalarca servis geldi”

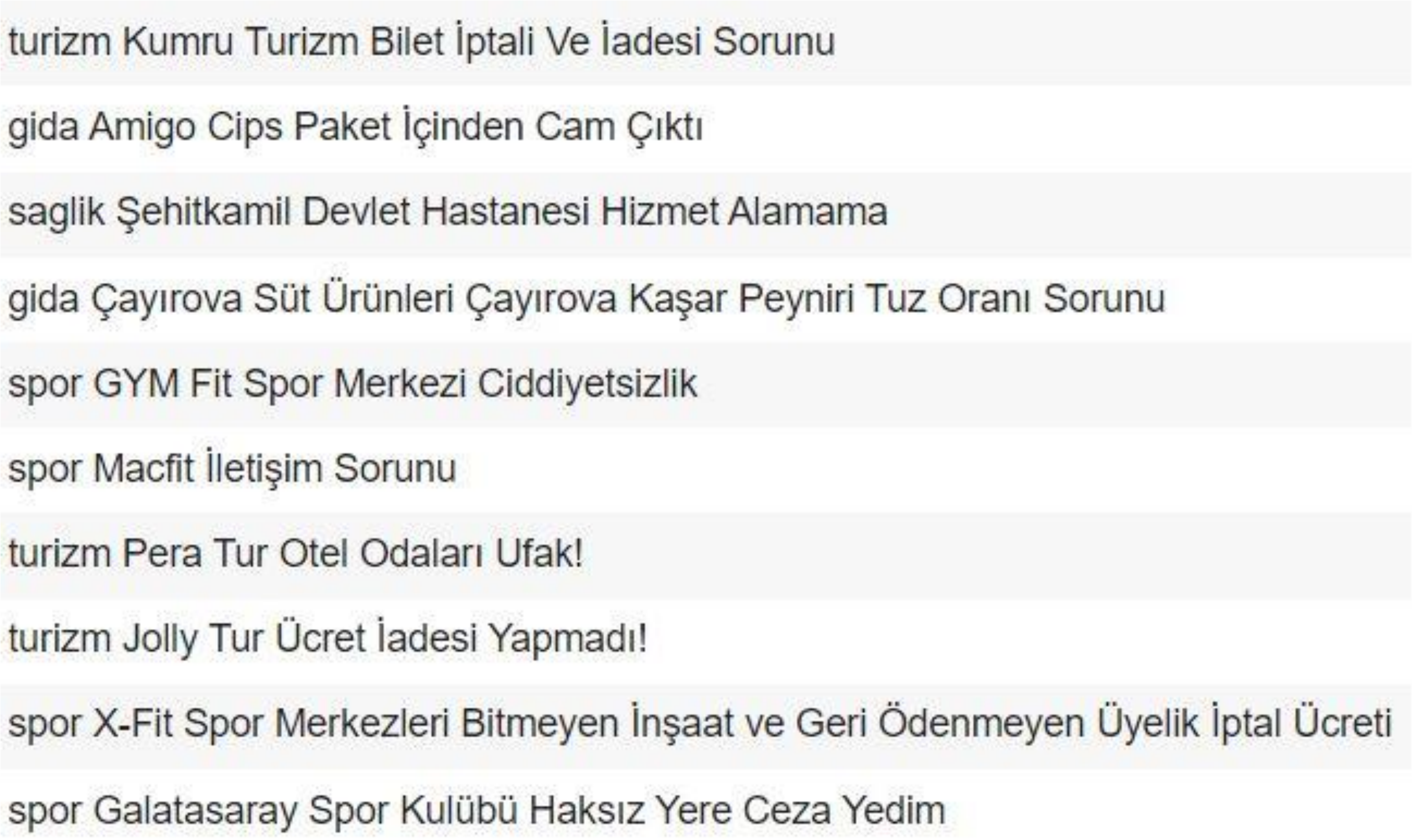
PROMPT GENERATOR

- To create automatic prompts for 5 topics.
- We can use first sentences (titles) of reviews.

Example from TC5:

”**Akbank Bilgim Dışında Vadeli Hesap Açma**,” ”Bilgim ve onayım olmadan vadeli nar hesabı açılmış nemalandırma adı altında benim vadesiz hesabımdaki paranın bir kısmı oraya aktarılmış. Vadeli hesabımdaki paranın vadesize aktarılmasını ve vadeli nar hesabımın kapatılmasını talep ediyorum.”

PROMPT GENERATOR DATASET



turizm Kumru Turizm Bilet İptali Ve İadesi Sorunu

gida Amigo Cips Paket İçinden Cam Çıktı

saglik Şehitkamil Devlet Hastanesi Hizmet Alamama

gida Çayırova Süt Ürünleri Çayırova Kaşar Peyniri Tuz Oranı Sorunu

spor GYM Fit Spor Merkezi Ciddiyetsizlik

spor Macfit İletişim Sorunu

turizm Pera Tur Otel Odaları Ufak!

turizm Jolly Tur Ücret İadesi Yapmadı!

spor X-Fit Spor Merkezleri Bitmeyen İnşaat ve Geri Ödenmeyen Üyelik İptal Ücreti

spor Galatasaray Spor Kulübü Haksız Yere Ceza Yedim

PROMPT GENERATOR

- When we want to create a Prompt Generator first we train the GPT architecture from scratch with this dataset but the results are not successful.

- Few data and short sentences

- Instead, we fine-tune the turkishReviews model we create, which we explained in the previous section. We use the trained weights and tokenizer of this model,

GENERATED PROMPTS

- Topic : turizm

Spa Club Fitness Termal Thermal
Otel Resort Termal Spa Merkezi
Hotel Resort Spa SPA Club Fitness Therma

- Topic : saglik

Doktor Polikliniği Doğum Kadın Doktoru
Acil Özel Üniversitesi Hastanesi Polikliniği
Devlet Eğitim Üniversitesi Araştırma Hastalıklar

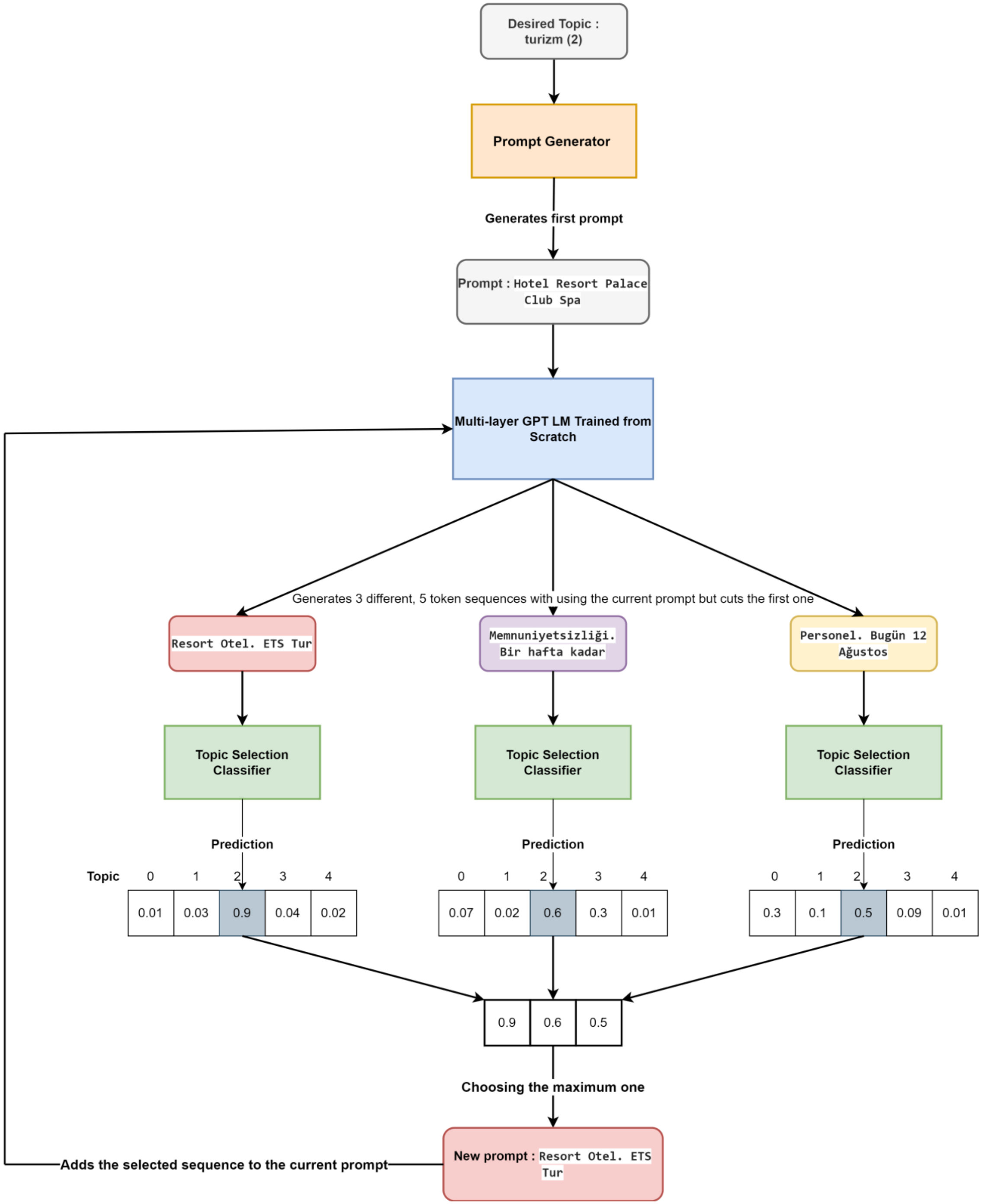
SAMPLING WITH TOPIC SELECTION CLASSIFIER

- Here we let the language model generate 5 sequential tokens instead of selecting one token at a time.

- Then we choose from the next sequence of these with usign the Topic Selection Classifier technique that we propose .

- Example :

- Resort Otel. ETS Tur
- Resort Otel. ETS Tur'dan 5 Gün Sonra rezervasyon



Desired Topic : turizm (2)

Prompt Generator

Generates first prompt

Prompt : Hotel Resort Palace Club Spa

Multi-layer GPT LM Trained from Scratch

Generates 3 different, 5 token sequences with using the current prompt but cuts the first one

Resort Otel. ETS Tur

Memnuniyetsizliği. Bir hafta kadar

Personel. Bugün 12 Ağustos

Topic Selection Classifier

Topic Selection Classifier

Topic Selection Classifier

Prediction

Topic	0	1	2	3	4
	0.01	0.03	0.9	0.04	0.02

Prediction

0	1	2	3	4
0.07	0.02	0.6	0.3	0.01

Prediction

0	1	2	3	4
0.3	0.1	0.5	0.09	0.01

0.9	0.6	0.5

Choosing the maximum one

New prompt : Resort Otel. ETS Tur

Adds the selected sequence to the current prompt

SAMPLING WITH TOPIC SELECTION CLASSIFIER

Generated Text	Desired Topic	Predicted Topic
Kredisi. 1 Nisan 2020’de krediye ait faiz oranı ile kredi kartım onaylandı ancak halen bir sonuç alamadım. Kredi kartımın ödemesi gerçekleştirmiyorlar. 2 aydır değerlendirme aşamasında yazıyor.	finans	finans
Kredi kartı ile ilgili mesaj yoluyla bildirim gelmedi üyelik için. Müşteri hizmetlerini arıyorum cevap alamıyorum. Ayrıca müşteri hizmetlerini açan yok açan yok bilgi alamıyorum	spor	finans

SAMPLING WITH TOPIC SELECTION CLASSIFIER

- An example with right topic but poor quality:

Desired topic : turizm

Jolly Tur den aldığımız tatil virüs nedeniyle seyahat sigortası iptal edindi dedim ama o gün iptal edin dedi bir de iptal edelim dedi 3 gün geçti iptal etmek istiyorum müşteri hizmetleri. Bugün müşteri hizmetleri olarak ulaşamıyorum

LM	Control Technique	BLEU	BERTScore (F1)	Classification Accuracy
Miniature GPT	Modifying Sequential Input with Topic	0.1046	0.39	0.66
Miniature GPT	Modifying Sequential Input with Keyword	0.1028	0.43	0.76
Miniature GPT	Sampling with Topic Selection Classifier	0.1010	0.44	0.88
Multi-layer GPT	Sampling with Topic Selection Classifier	0.1213	0.48	0.91

GENERATED REVIEWS WITH MINIATURE GPT

- Modifying Sequential Input with Topic

gelen sifre gonderilemiyor diyor oldugum kredi karti tarihinde tl tutarinda bir turlu zorlugu bacakta emboli teshisi konuldu ibaresi cikiyor ben boyle bir suru insan

- Modifying Sequential Input with Keywords

hukuk burosu hatali ayinda alinan urunun parasini ayinda kredi karti talebinde bulunmustum bana gelen var ancak son care bulamadilar inanilmaz derecede sikayetciyiz

MINITAUTURE GPT VS MULTI-LAYER GPT

- Sampling with Topic Selection Classifier (Miniature GPT)

kabul disindan param gundur suredir akbank musterisiyim ve temsilciler kendilerinden yasca kisa sure once tl odeyerek telefon ettigimde ufff dair bir sey demiyor

- Sampling with Topic Selection Classifier (Multi-layer GPT)

Bireysel destek kredisine başvurdum. Pandemi destek kredisine. Bir süre sonra başvuran destek kredisi alamadım. Hala hiçbir cevap alamadım. Nasıl bir mesaj geliyor.

REFERENCES

1. Topical Language Generation using Transformers, Rohola Zandie and Mohammad H. Mahoor, 2021.
2. T. Noraset, C. Liang, L. Birnbaum, and D. Downey, “Definition modeling: Learning to define word embeddings in natural language,” 31st AAAI Conference on Artificial Intelligence, AAAI 2017, pp. 3259–3266, 2017.
3. K. Zhou, S. Prabhumoye, and A. W. Black, “A dataset for document grounded conversations,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 708–713, 2020.
4. S. Prabhumoye, C. Quirk, and M. Galley, “Towards content transfer through grounded text generation,” NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 2622–2632, 2019
5. V. Harrison, L. Reed, S. Oraby, and M. Walker, “Maximizing stylistic control and semantic accuracy in NLG: Personality variation and discourse contrast,” DSNNLG
6. K. Chandu, S. Prabhumoye, R. Salakhutdinov, and A. W. Black, “"My Way of Telling a Story": Persona based Grounded Story Generation,” pp. 11–21, 2019.
7. C. Caglayan and M. Karakaya, ”Topic-Controlled Text Generation,” 6th International Conference on Computer Science and Engineering (UBMK), pp. 533-536, Sep. 2021.
8. https://www.kaggle.com/datasets/savasy/multiclass-classification-data-for-turkish-tc32
9. Attention Is All You Need , Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, 2017.
10. https://huggingface.co/kmkarakaya/turkishReviews-ds

Thank you for listening.

Cansen Çağlayan