

ABSCHNITT 6

# NATURAL LANGUAGE PROCESSING

TEXTDATEN AUTOMATISIERT VERARBEITEN

A close-up photograph of a white humanoid robot's face. The robot has large, black, almond-shaped eyes with a bright blue glow from within. It has a small, thin black smile. On the right side of its face, there is a circular speaker grille. The background is blurred, showing some yellow and grey colors.

NLP KÖNNTE DER SCHLÜSSEL ZU ECHTER  
KÜNSTLICHER INTELLIGENZ SEIN.

# AM ANFANG STAND DAS GEORGETOWN IMB EXPERIMENT (1954).



Die Überzeugung:  
„Binnen drei Jahren wird maschinelles  
Übersetzen ein gelöstes Problem sein.“

# SPOILER: MASCHINELLES ÜBERSETZEN IST HEUTE NOCH IMMER NICHT GELÖST.

Sprache ist generativ.

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, ...

Sprache kann korrekt aber inhaltslos sein.

Sprache ist mehrdeutig.

Ich sitze bei der Bank.

Nachts ist es kälter als draußen.

# SPRACHEN SIND SEHR UNTERSCHIEDLICH.

Chinesisch

tóngxuémen

同学们

analytisch

logosyllabisch

Deutsch

Kommilitonen

synthetisch

alphabetisch

# NLP IST EXTREM RECHENINTENSIV.

Zum Modellieren der schieren Vielfalt braucht man riesige Datensätze.

Menschliche Sprache zu modellieren ist komplex. Wie sich Sprache im Menschen entwickelt, ist selbst noch nicht vollständig verstanden.

Viele komplexe Modelle sind denkbar, aber die rechnerischen Ressourcen reichen nicht aus.

Sprache beinhaltet stilistische Komponenten wie Ironie.

---

# **WELCHE KATEGORIEN UMFASST NLP?**

---

# NLP BEFASST SICH MIT GESCHRIEBENER UND GESPROCHENER SPRACHE.

## **Speech recognition.**

Das Erkennen gesprochener Sprache. Übersetzen in Text oder ein für Maschinen verständliches Format. Beispiel: Siri.

## **Natural language understanding.**

Auch: machine reading. Maschinelles Erfassen der Bedeutung von geschriebener oder gesprochener Sprache.

## **Natural language generation.**

Maschinelles Erzeugen von geschriebener oder gesprochener Sprache. Beispiel: Maschinelles Beantworten von Fragen.

---

# TYPISCHE AUFGABEN IM NLP SIND:

PoS tagging

Named entity recognition

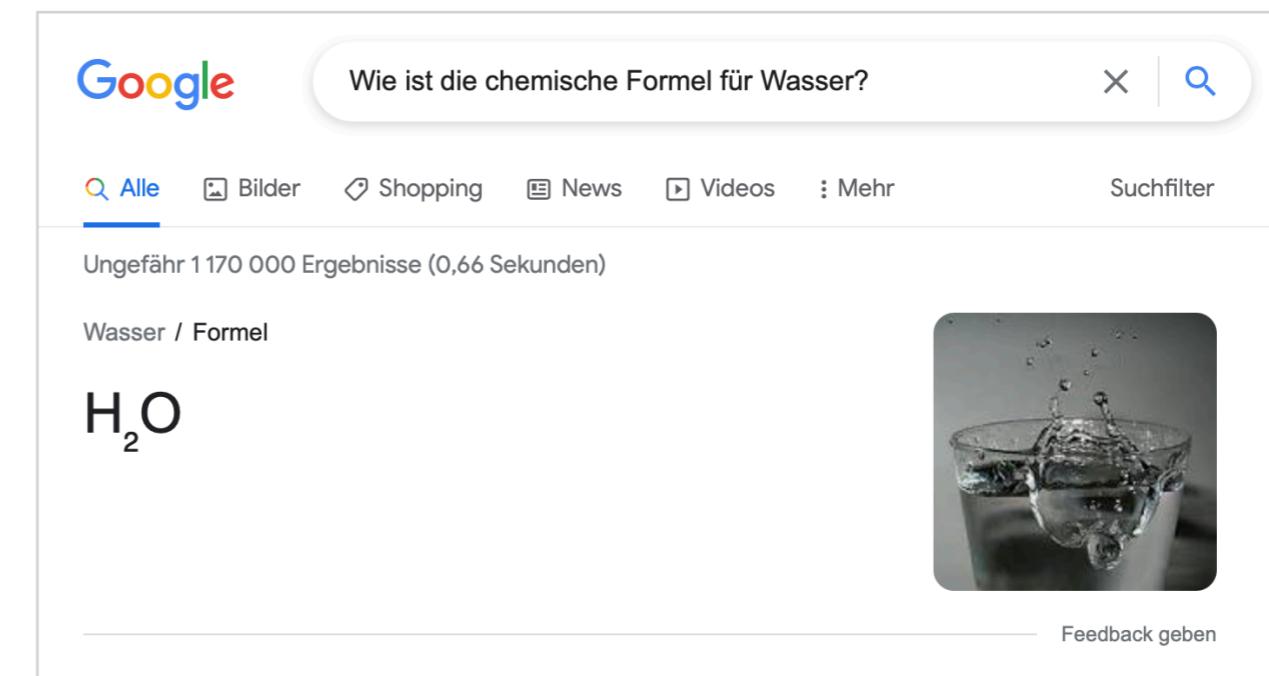
Semantic similarity

Document classification

Text summarization

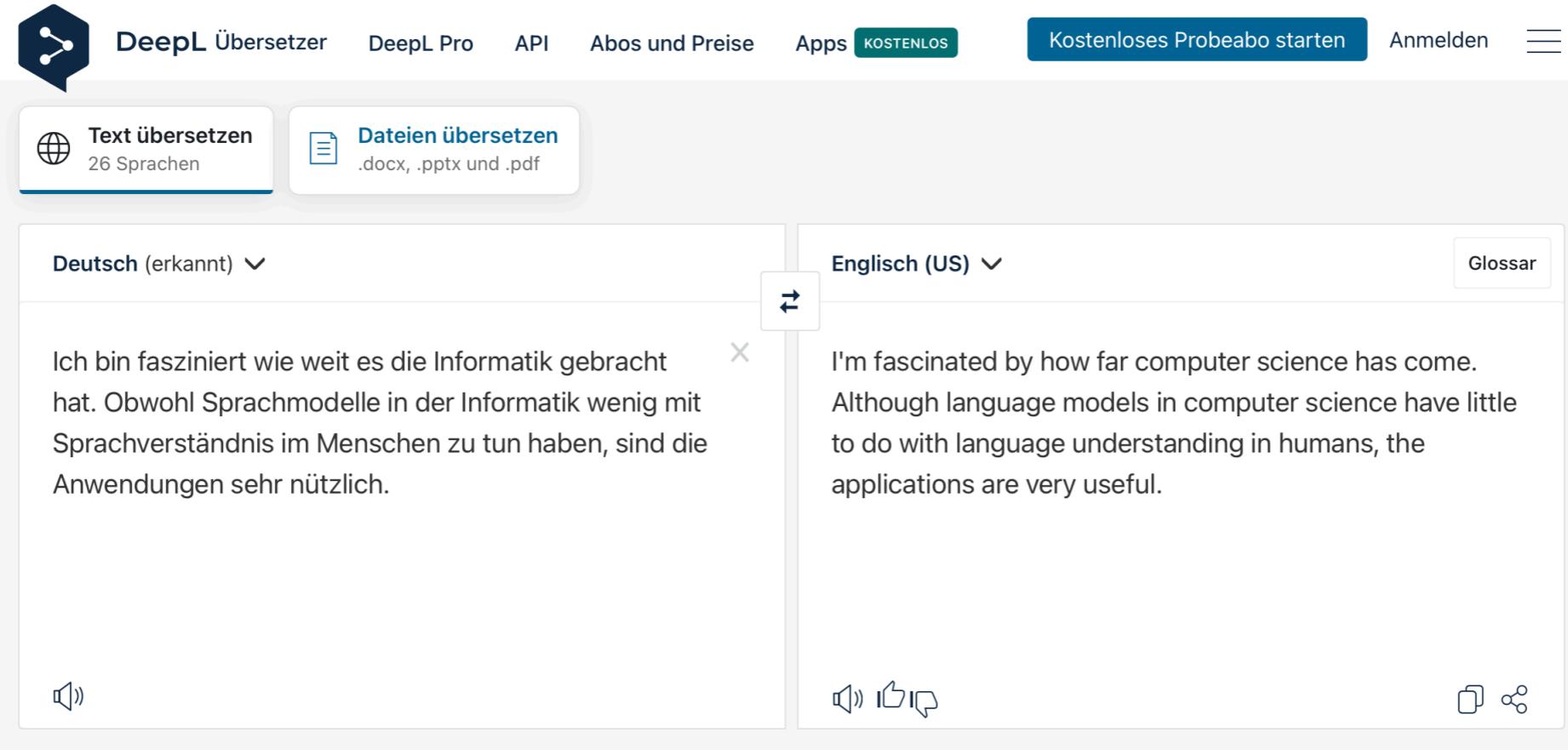
Question answering

Sentiment analysis



Question Answering

# EINIGE ANWENDUNGSBEISPIELE: MASCHINELLES ÜBERSETZEN.



The screenshot shows the DeepL Translator interface. At the top, there's a navigation bar with links for "DeepL Übersetzer", "DeepL Pro", "API", "Abos und Preise", "Apps" (with a "KOSTENLOS" button), "Kostenloses Probeabo starten", "Anmelden", and a menu icon. Below the navigation bar are two main translation options: "Text übersetzen" (26 Sprachen) and "Dateien übersetzen" (.docx, .pptx und .pdf). The main area has two language dropdowns: "Deutsch (erkannt)" on the left and "Englisch (US)" on the right. A central double-headed arrow icon indicates the direction of translation. On the left, the German input text is:

Ich bin fasziniert wie weit es die Informatik gebracht hat. Obwohl Sprachmodelle in der Informatik wenig mit Sprachverständnis im Menschen zu tun haben, sind die Anwendungen sehr nützlich.

On the right, the English output is:

I'm fascinated by how far computer science has come. Although language models in computer science have little to do with language understanding in humans, the applications are very useful.

At the bottom of each text area are small icons for audio playback, like/dislike, and sharing.

# EINIGE ANWENDUNGSBEISPIELE: CHATBOTS.

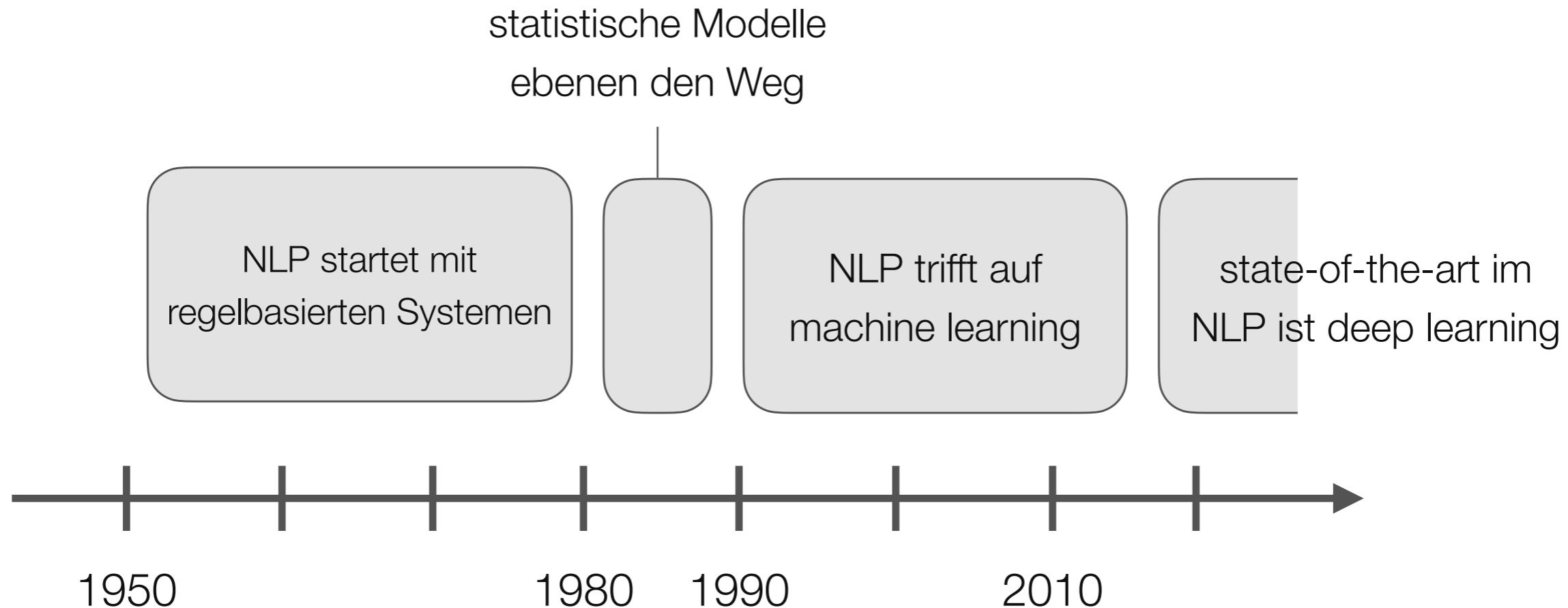
The screenshot shows a web page for Wien Energie. At the top, there is a navigation bar with the Wien Energie logo, a search bar, and links for 'Produkte' (Products), 'Erleben' (Experience), 'Strom', 'Erdgas', 'Wärme', 'Sonnenenergie', 'E-Mobilität', 'Internet.TV.Telefonie', and 'Mobilfunk'. A prominent orange banner on the left side contains an information icon and the text: 'Coronavirus: WIR SIND WEITER FÜR SIE DA!'. Below this, there is a paragraph about service access during the pandemic. On the right side, a chatbot window titled 'BotTina' is open. It features a cartoon illustration of a woman with orange hair. The conversation starts with 'Willkommen bei Wien Energie!', followed by 'Hallo, ich bin BotTina.', and a message asking for permission to store and analyze user data. At the bottom of the chat window, there are buttons for 'Ja', 'Infos Datenschutz', and 'Nein', along with a text input field and a microphone icon. The URL <https://www.wienenergie.at> is visible at the bottom of the page.

---

**NLP HAT EINEN WEITEN WEG  
HINTER SICH**

---

# HEUTE HERRSCHEN IM NLP DEEP LEARNING MODELLE VOR.



---

1990 - 2015

# NLP & MACHINE LEARNING

---

# MACHINE LEARNING MODELLE VERSTEHEN AUSSCHLIESSLICH ZAHLEN.

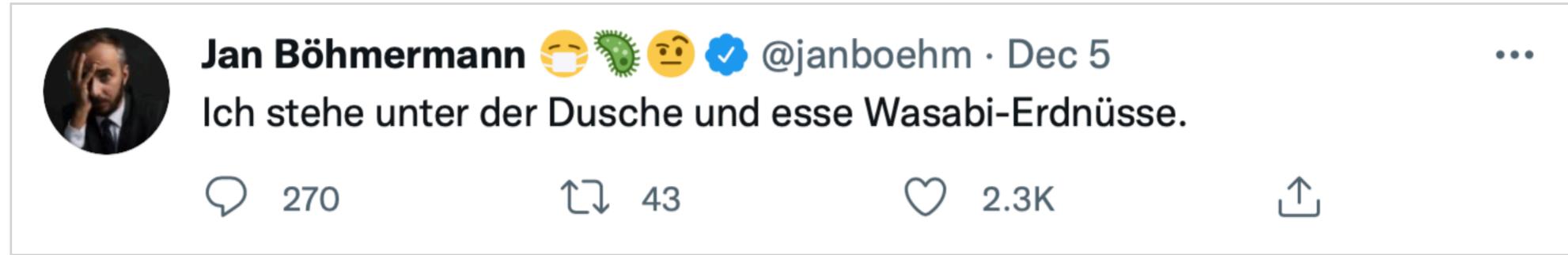
Bevor wir unsere Textdaten einem ausgewählten Machine Learning Modell präsentieren, müssen wir sie in die geeignete Form bringen.

Stellen wir uns vor wir haben eine Reihe Social Media Texte - zum Beispiel Tweets - und möchten diese in Sentiment Kategorien klassifizieren (positiv, neutral, negativ) ...

---

# **TEXT IN ZAHLEN ÜBERSETZEN: DAS PREPROCESSING**

# (1/6) NORMALIZATION



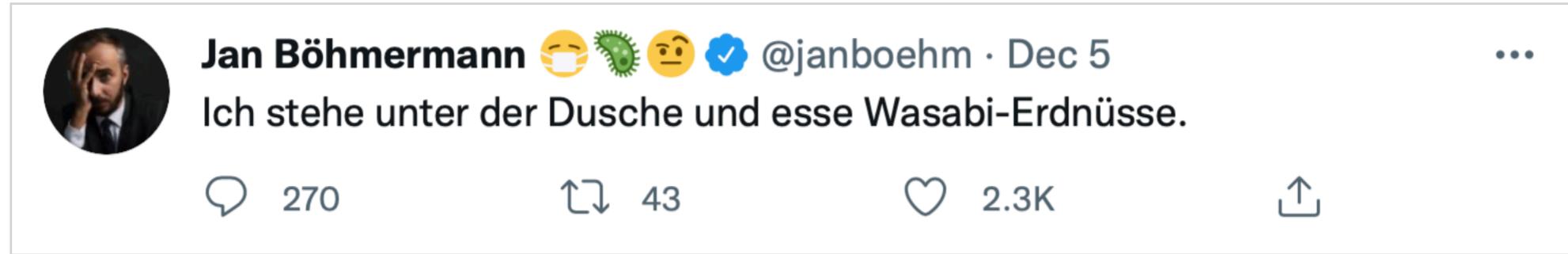
A screenshot of a Twitter post from Jan Böhmermann (@janboehm). The post was made on December 5 and contains the following text: "Ich stehe unter der Dusche und esse Wasabi-Erdnüsse." The tweet includes several emojis: a face with a mask, a green virus-like emoji, a face with a headband, and a blue checkmark. The post has 270 replies, 43 retweets, and 2.3K likes. There is also a share icon.

ich stehe unter der dusche und esse wasabierdnuesse

„Wasabierdnüsse“

„Wasabi-Erdnuesse“

## (2/6) TOKENIZATION



A screenshot of a Twitter post from Jan Böhmermann (@janboehm). The post was made on December 5 and contains the following text: "Ich stehe unter der Dusche und esse Wasabi-Erdnüsse." The tweet includes three emojis: a face with a mask, a green virus-like cell, and a face with a headband. It has 270 replies, 43 retweets, and 2.3K likes. There are three dots at the end of the text and a share icon.

[,ich‘, ,stehe‘, ,unter‘, ,der‘, ,dusche‘, ,und‘, ,esse‘, ,wasabi‘, ,erdnuesse‘]

don't > [don, t] [do, n't] [do not] ?

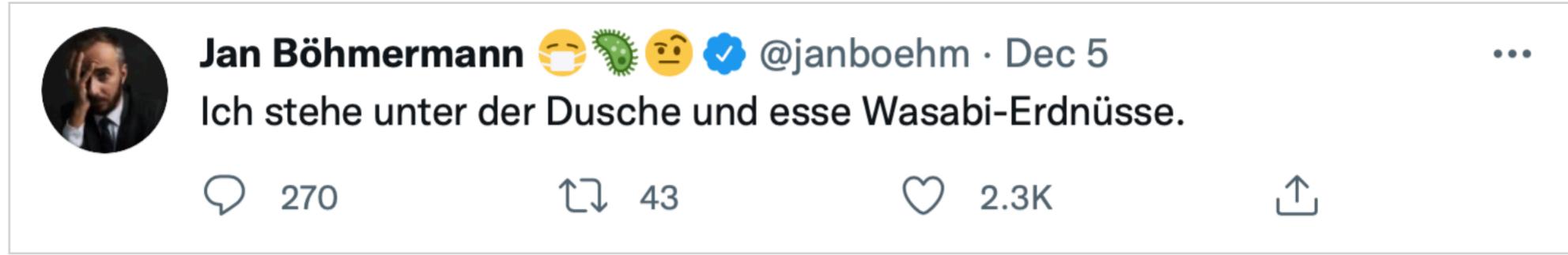
# (3/6) REMOVING STOPWORDS

```
✓ [1] import nltk
0 s   nltk.download('stopwords')
      from nltk.corpus import stopwords

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!

✓ ⏎ stopwords.words('german')
0 s
  ↳ 'sie',
    'ihnen',
    'sind',
    'so',
    'solche',
    'solchem',
    'solchen',
    'solcher',
    'solches',
    'soll',
    'sollte',
    'sondern',
    'sonst',
    'über',
    'um',
    'und',
```

## (3/6) REMOVING STOPWORDS



A screenshot of a Twitter post from Jan Böhmermann (@janboehm). The post includes a profile picture of a man with a beard, three emojis (mask, virus, face with head-bust), and a blue checkmark. The text reads: "Ich stehe unter der Dusche und esse Wasabi-Erdnüsse." Below the tweet are engagement metrics: 270 replies, 43 retweets, 2.3K likes, and a share icon.

[,stehe‘, ,dusche‘, ,esse‘, ,wasabi‘, ,erdnuesse‘]

# (4/6) STEMMING



**Jan Böhmermann** 😷🦠😉 ✅ @janboehm · Dec 5

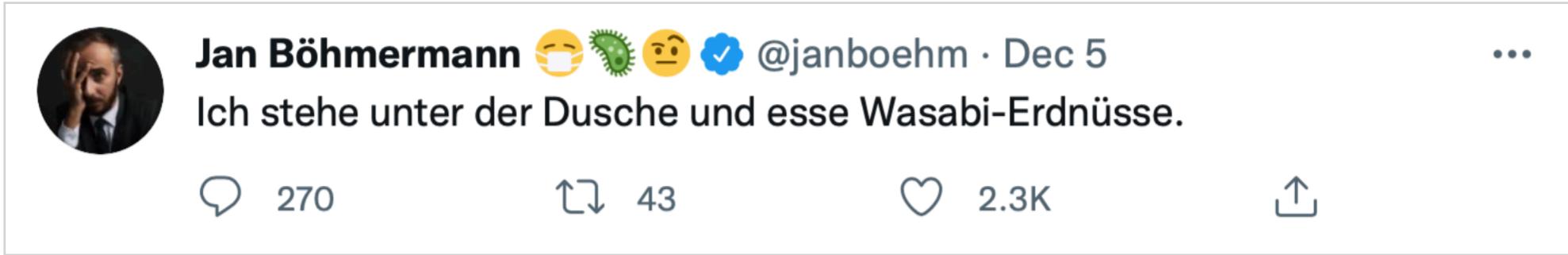
Ich stehe unter der Dusche und esse Wasabi-Erdnüsse.

270 43 2.3K

...

[,steh‘, ,dusch‘, ,ess‘, ,wasabi‘, ,erdnuess‘]

## (5/6) LEMMATIZATION



A screenshot of a Twitter post from Jan Böhmermann (@janboehm). The post includes a profile picture of a man with a beard, three emojis (mask, virus, face with mask), a blue checkmark, the handle @janboehm, the date Dec 5, and the text "Ich stehe unter der Dusche und esse Wasabi-Erdnüsse." Below the tweet are engagement metrics: 270 replies, 43 retweets, 2.3K likes, and a share icon.

[,stehen‘, ,dusche‘, ,essen‘, ,wasabi‘, ,erdnuesse‘]

# (6/6) N-GRAMS BILDEN



**Jan Böhmermann** 😷🦠🤔 ✅ @janboehm · Dec 5

Ich stehe unter der Dusche und esse Wasabi-Erdnüsse.

270 43 2.3K

...

[,stehen\_dusche‘, ,dusche\_essen‘, ,essen\_wasabi‘, ,wasabi\_erdnuesse‘]

# DIE PREPROCESSING SCHRITTE SIND ABHÄNGIG VON DER ZU LÖSENDE AUFGABE!

Für unsere vorliegende Aufgabe (Sentiment Analyse von Tweets) könnten zum Beispiel noch folgende Schritte sinnvoll sein:

- URLs, Hashtags und User-tags entfernen
- Emojis durch Worte ersetzen, die deren Emotionen beschreiben
- eine Rechtschreibkorrektur durchführen
- Abkürzungen ausschreiben

# EINE EINZIGE REPRÄSENTATION FÜR EIN DOKUMENT FINDEN

# EIN DENKBAR EINFACHES KONZEPT: ,BAG OF WORDS'.



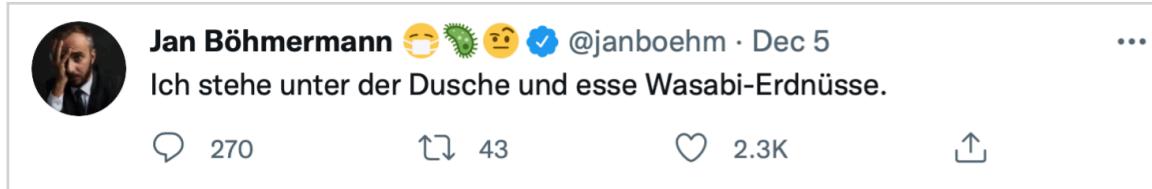
[,stehen‘, ,dusche‘, ,essen‘, ,wasabi‘, ,erdnuesse‘]

[,richard‘, ,david‘, ,precht‘, ,eitelkeit‘, ,impfen‘]

$$t1 = [1, 2, 3, 4, 5]$$

$$t2 = [6, 7, 8, 9, 10]$$

# EIN VERWANDTER DES BAG OF WORDS: DIE DOCUMENT-TERM-MATRIX.



[,stehen‘, ,dusche‘, ,essen‘, ,wasabi‘, ,erdnuesse‘]

[,richard‘, ,david‘, ,precht‘, ,eitelkeit‘, ,impfen‘]

	1	2	3	4	5	6	7	8	9	10
t1	1	1	1	1	1	0	0	0	0	0
t2	0	0	0	0	0	1	1	1	1	1

# DIE VERBESSERTE DOCUMENT-TERM-MATRIX: TF-IDF

TF-IDF = Term-Frequency Inverse-Document-Frequency

$TF(w_i, d_j) = \# w_i \text{ befindet sich in } d_j / \text{gesamt } \# \text{ aller Worte in } d_j$

$IDF(w_i, D_c) = \log(N/n_i)$

$TF-IDF = TF * IDF$

$w_i$  ... Wort i

$d_j$  ... Dokument j

$D_c$  ... alle Dokumente im Korpus

N ... Anzahl aller Dokumente

$n_i$  ... Dokumente, die das Wort  $w_i$  beinhalten

# UNSER TEXT IST JETZT BEREIT FÜR EINEN MACHINE LEARNING ALGORITHMUS!

Für unsere Sentiment Analyse von Tweets, genauer gesagt die Kategorisierung von Tweets in drei diskrete Klassen (positiv, neutral, negativ), könnten wir also bekannte Klassifizierungsalgorithmen verwenden, z.B. die logistische Regression oder SVMs.

---

2015 - HEUTE

# DEEP LEARNING FÜR NLP

---

# **FROM SPARSE TO DENSE: FORTGESCHRITTENE REPRÄSENTATION VON DOKUMENTEN**

---

# EINE NEUE ART SPRACHE ZU REPRÄSENTIEREN: WORD EMBEDDINGS

## Efficient Estimation of Word Representations in Vector Space

**Tomas Mikolov**

Google Inc., Mountain View, CA  
tmikolov@google.com

**Kai Chen**

Google Inc., Mountain View, CA  
kaichen@google.com

**Greg Corrado**

Google Inc., Mountain View, CA  
gcorrado@google.com

**Jeffrey Dean**

Google Inc., Mountain View, CA  
jeff@google.com

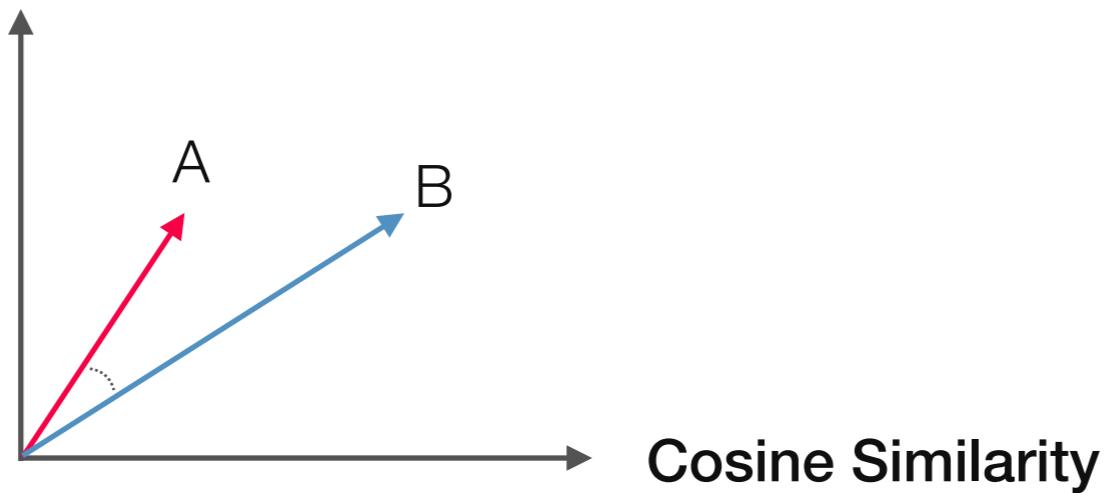
### Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

# WIE WERDEN WORD EMBEDDINGS ERZIELT?

**Distributional Hypothesis:** Worte, die in ähnlichen Kontexten auftreten, sind sich auch semantisch ähnlich.

„Heute ist ein toller Tag.“ // „Heute ist ein großartiger Tag.“



# WIE WERDEN WORD EMBEDDINGS ERZIELT? DIE COMMON BAG OF WORDS TECHNIK.

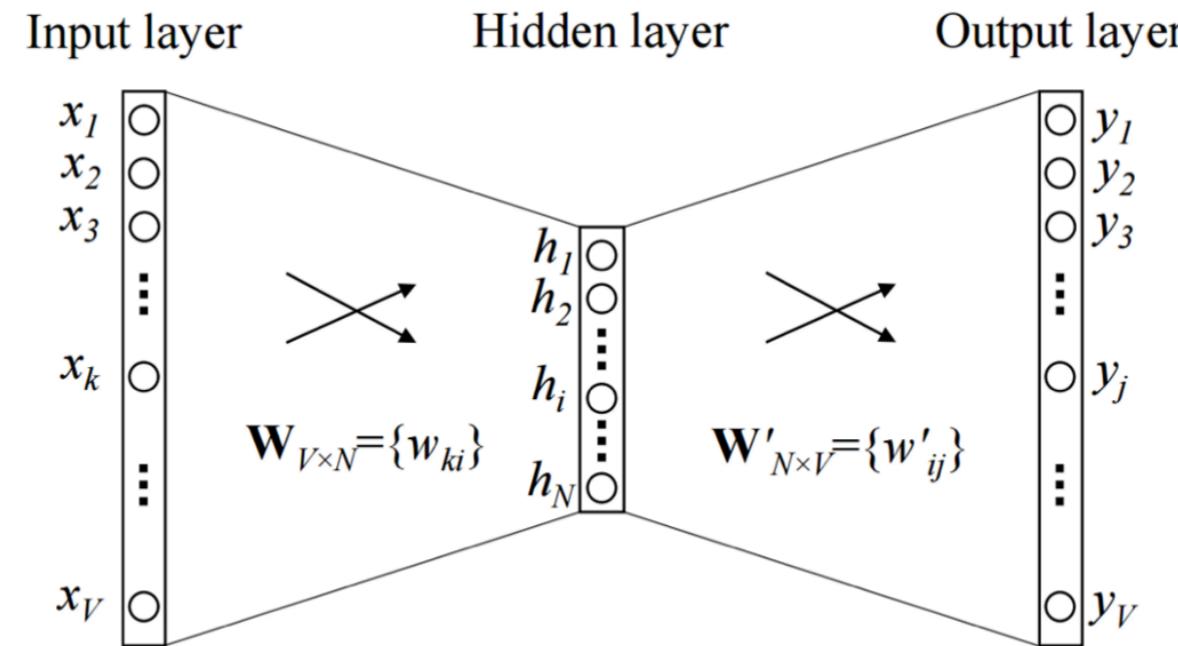
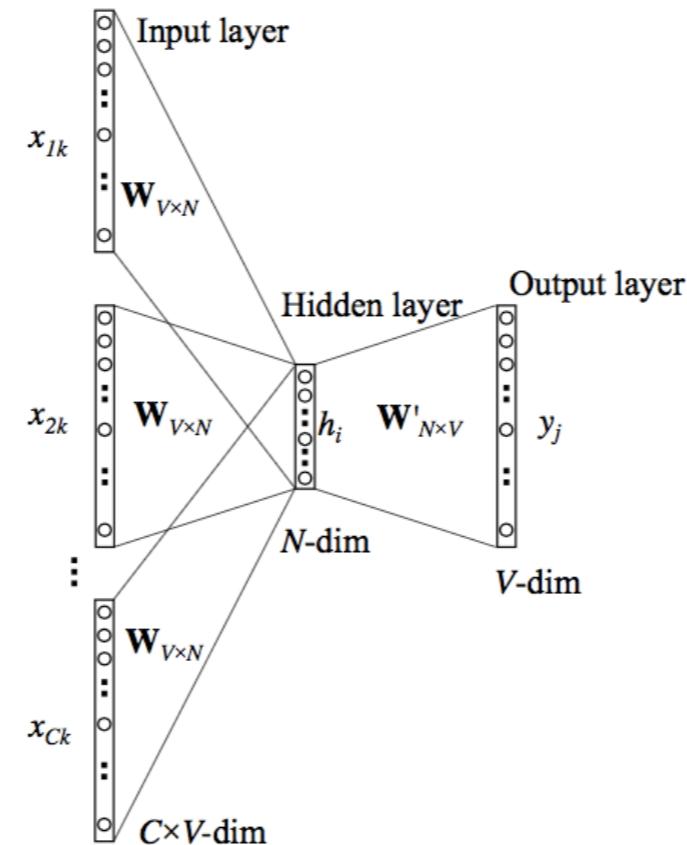
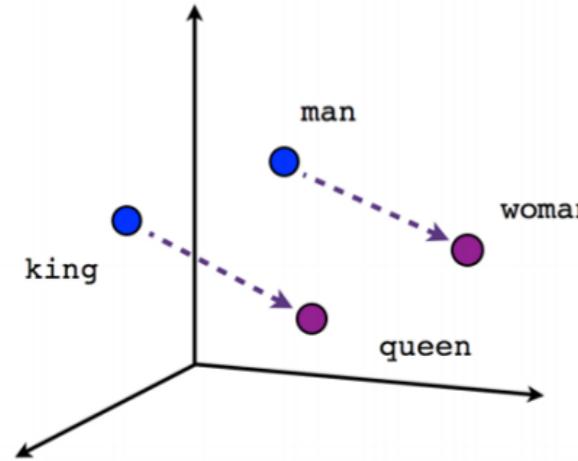


Figure 1: A simple CBOW model with only one word in the context

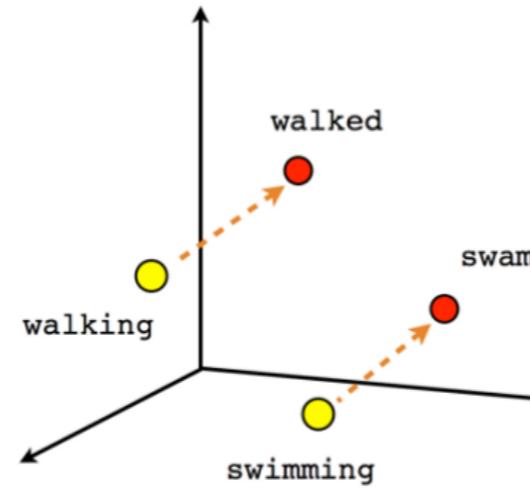
# WIE WERDEN WORD EMBEDDINGS ERZIELT? DIE SKIP GRAM METHODE.



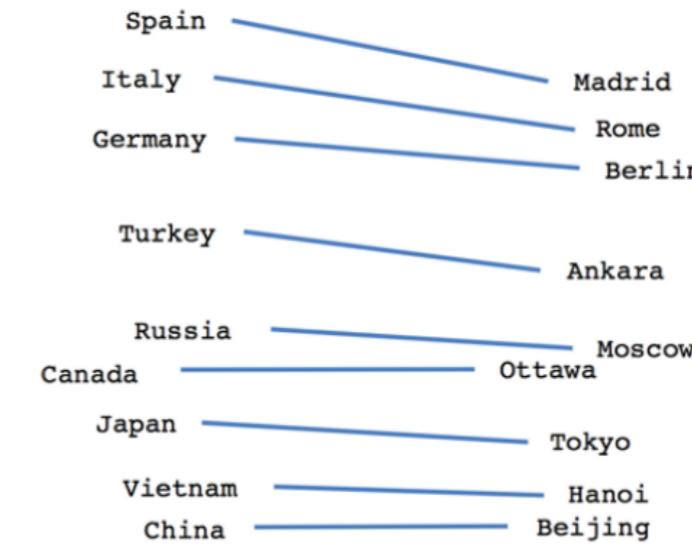
# WORD EMBEDDINGS SPIEGELN SEMANTISCHE VERWANDTSCHAFT WIEDER!



Male-Female



Verb tense



Country-Capital

# ACHTUNG: WORD EMBEDDINGS SPIEGELN AUCH MENSCHLICHEN BIAS WIEDER!

Man is to Computer Programmer as Woman is to Homemaker?

Debiasing Word Embeddings

Tolga Bolukbasi<sup>1</sup>, Kai-Wei Chang<sup>2</sup>, James Zou<sup>2</sup>, Venkatesh Saligrama<sup>1,2</sup>, Adam Kalai<sup>2</sup>

<sup>1</sup>Boston University, 8 Saint Mary's Street, Boston, MA

<sup>2</sup>Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com



# NEBEN MIKOLOV'S WORD2VEC EXISTIEREN WEITERE WORD EMBEDDING TECHNIKEN.

**GloVe**: von Standford Forschenden entwickelt; beachtet neben lokalen Abhängigkeiten von Wörtern auch deren globale Abhängigkeiten

**Elmo**: vom Allen Institute of AI entwickelt; kreiert unterschiedliche Vektoren für Polyseme, d.h. gleiche Worte mit unterschiedlicher Bedeutung („Fliege“)

**fasttext**: von Facebook entwickelt; ist ein multilinguales word embedding

# VON WORD EMBEDDINGS ZU DOKUMENTEN REPRÄSENTATIONEN.

Um von Word Embeddings zu Repräsentationen von Sätzen oder Texten zu gelangen, gibt es verschiedene Techniken. Eine Technik wäre zum Beispiel den Mittelpunkt aller Worte eines Dokumentes im Word Embedding Raum zu bestimmen.

---

DER NEUSTE SCHREI  
**TRANSFORMERS: ENCODER-  
DECODER ARCHITECTURES**

---

# GOOGLE DID IT AGAIN.

## Attention Is All You Need

**Ashish Vaswani\***  
Google Brain  
[avaswani@google.com](mailto:avaswani@google.com)

**Noam Shazeer\***  
Google Brain  
[noam@google.com](mailto:noam@google.com)

**Niki Parmar\***  
Google Research  
[nikip@google.com](mailto:nikip@google.com)

**Jakob Uszkoreit\***  
Google Research  
[usz@google.com](mailto:usz@google.com)

**Llion Jones\***  
Google Research  
[llion@google.com](mailto:llion@google.com)

**Aidan N. Gomez\* †**  
University of Toronto  
[aidan@cs.toronto.edu](mailto:aidan@cs.toronto.edu)

**Lukasz Kaiser\***  
Google Brain  
[lukaszkaiser@google.com](mailto:lukaszkaiser@google.com)

**Illia Polosukhin\* ‡**  
[illia.polosukhin@gmail.com](mailto:illia.polosukhin@gmail.com)

# Encoder

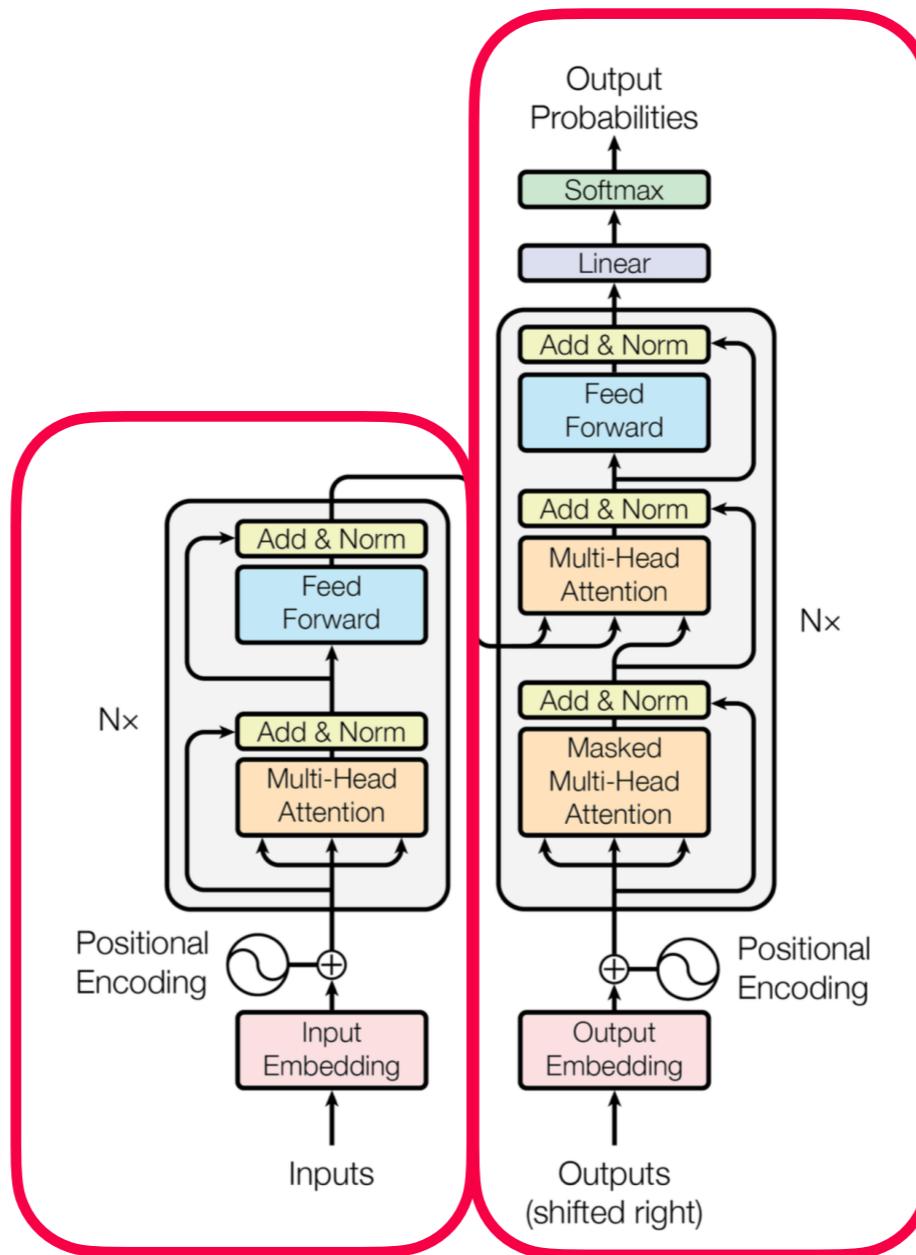


Figure 1: The Transformer - model architecture.

# Decoder

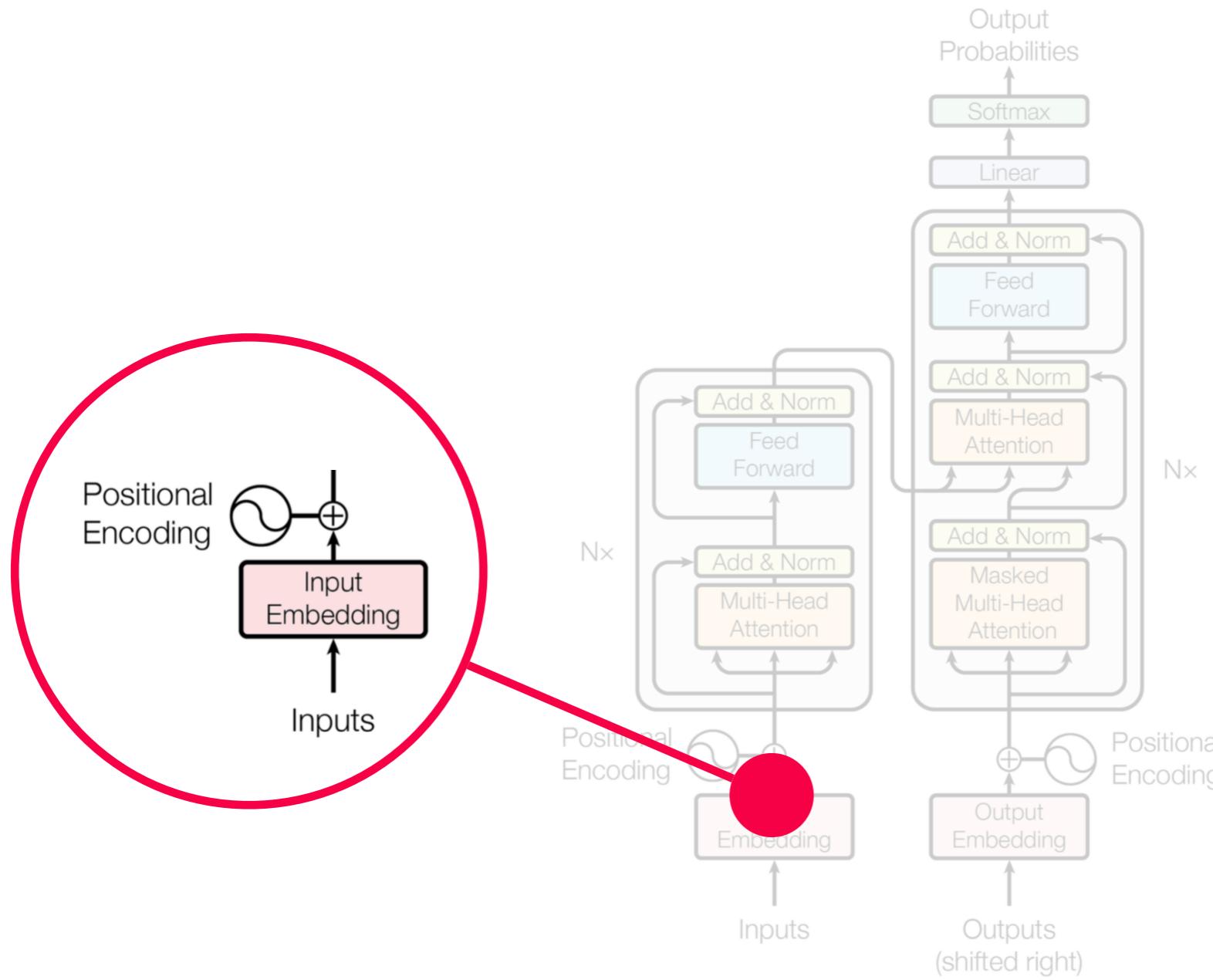


Figure 1: The Transformer - model architecture.

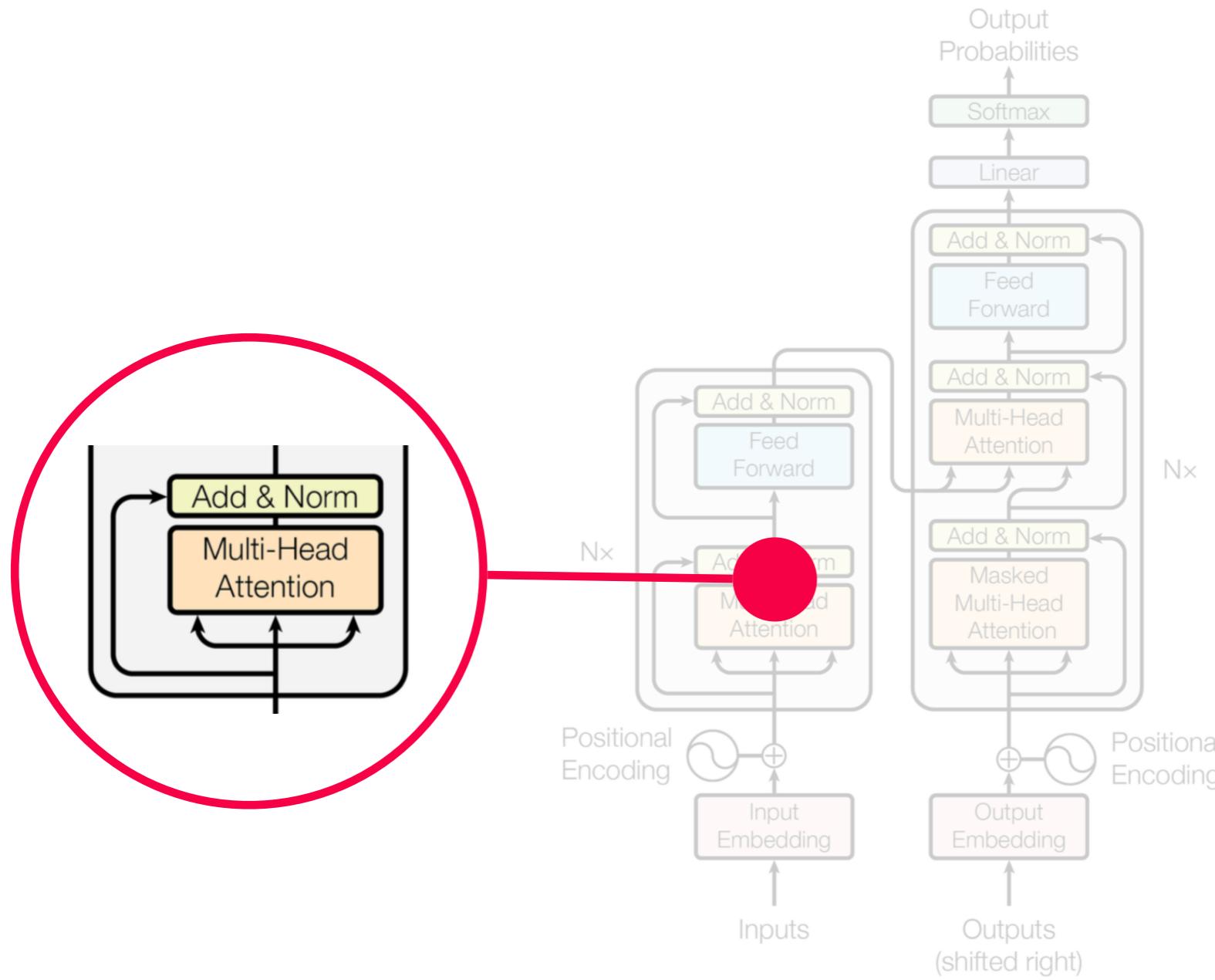


Figure 1: The Transformer - model architecture.

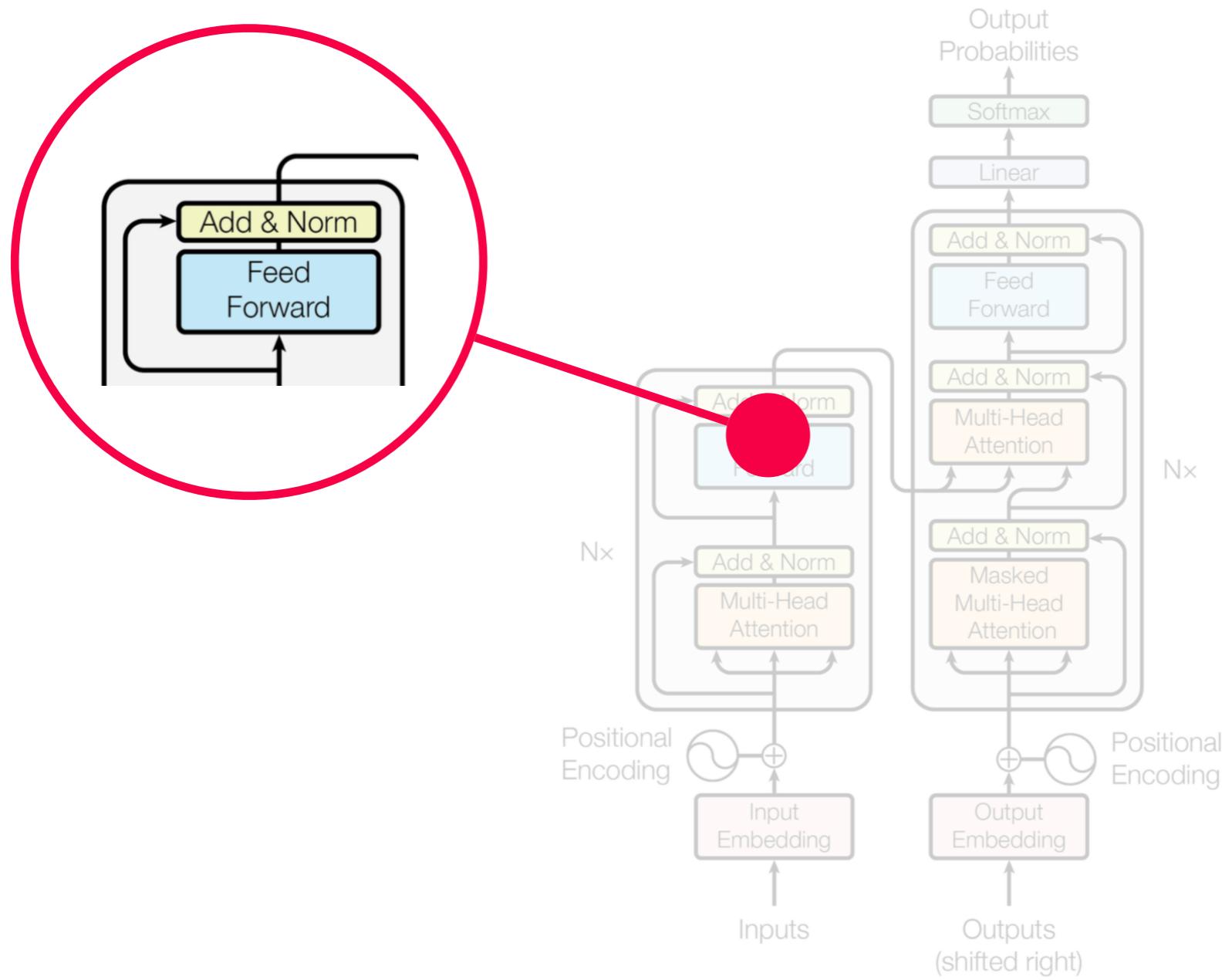


Figure 1: The Transformer - model architecture.

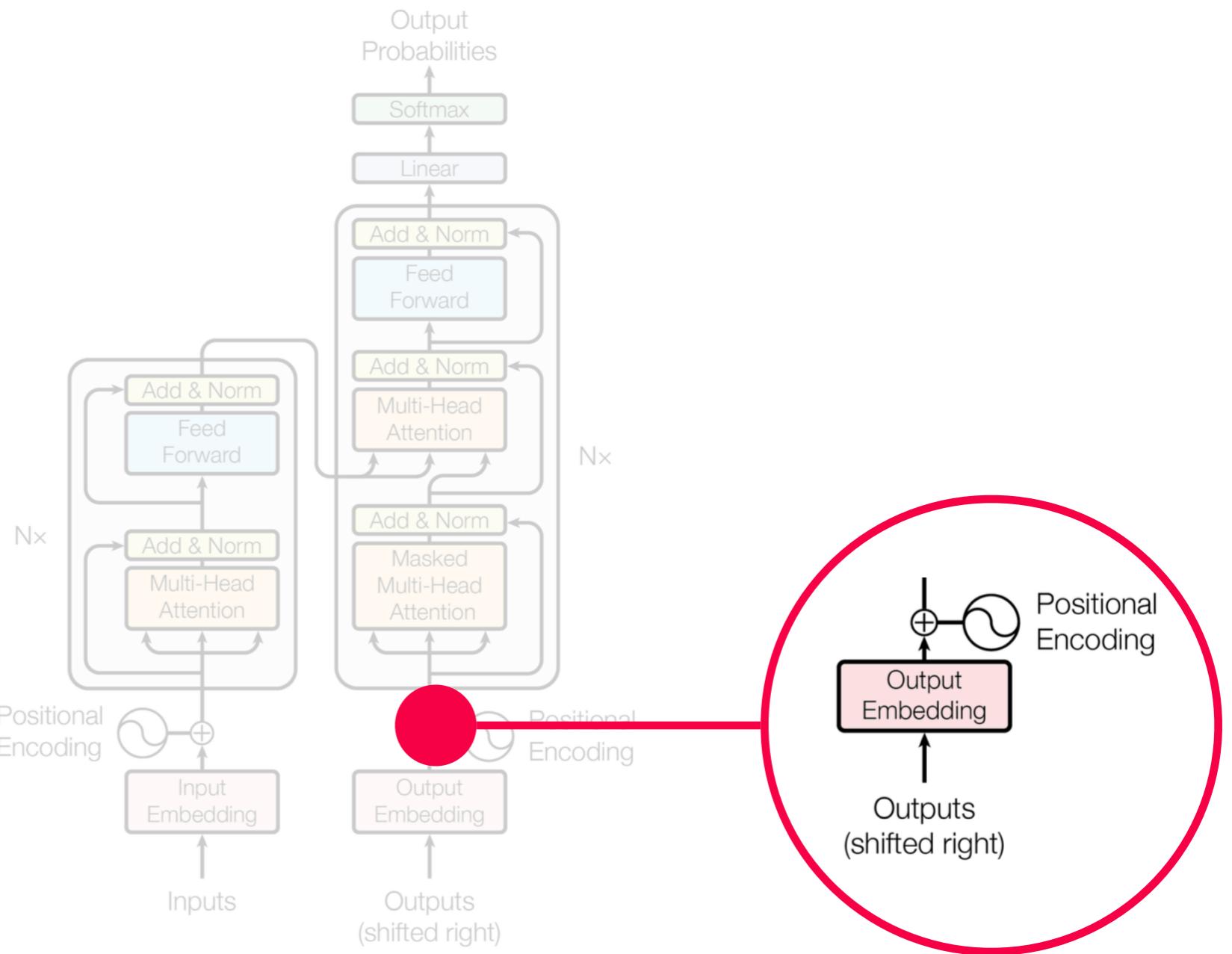


Figure 1: The Transformer - model architecture.

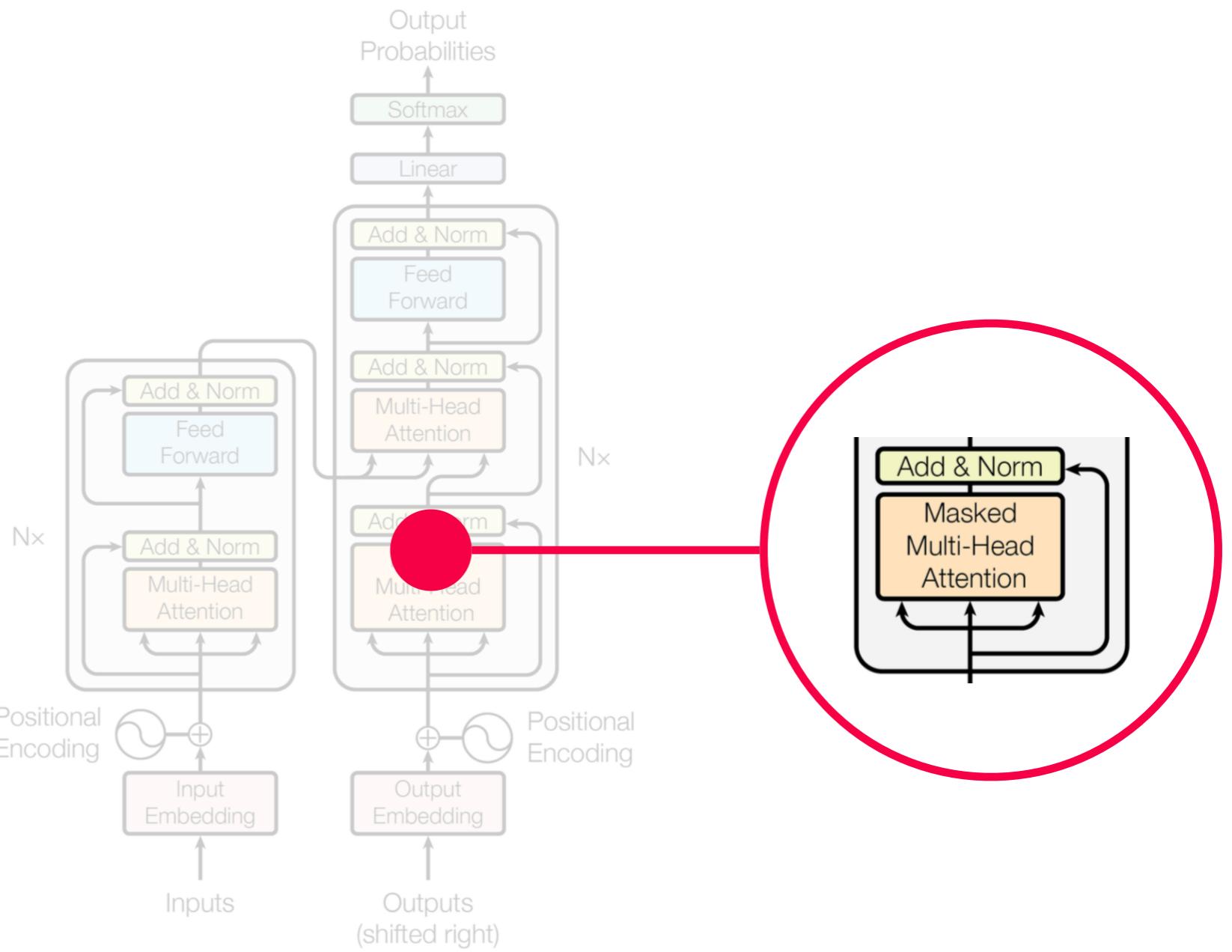


Figure 1: The Transformer - model architecture.

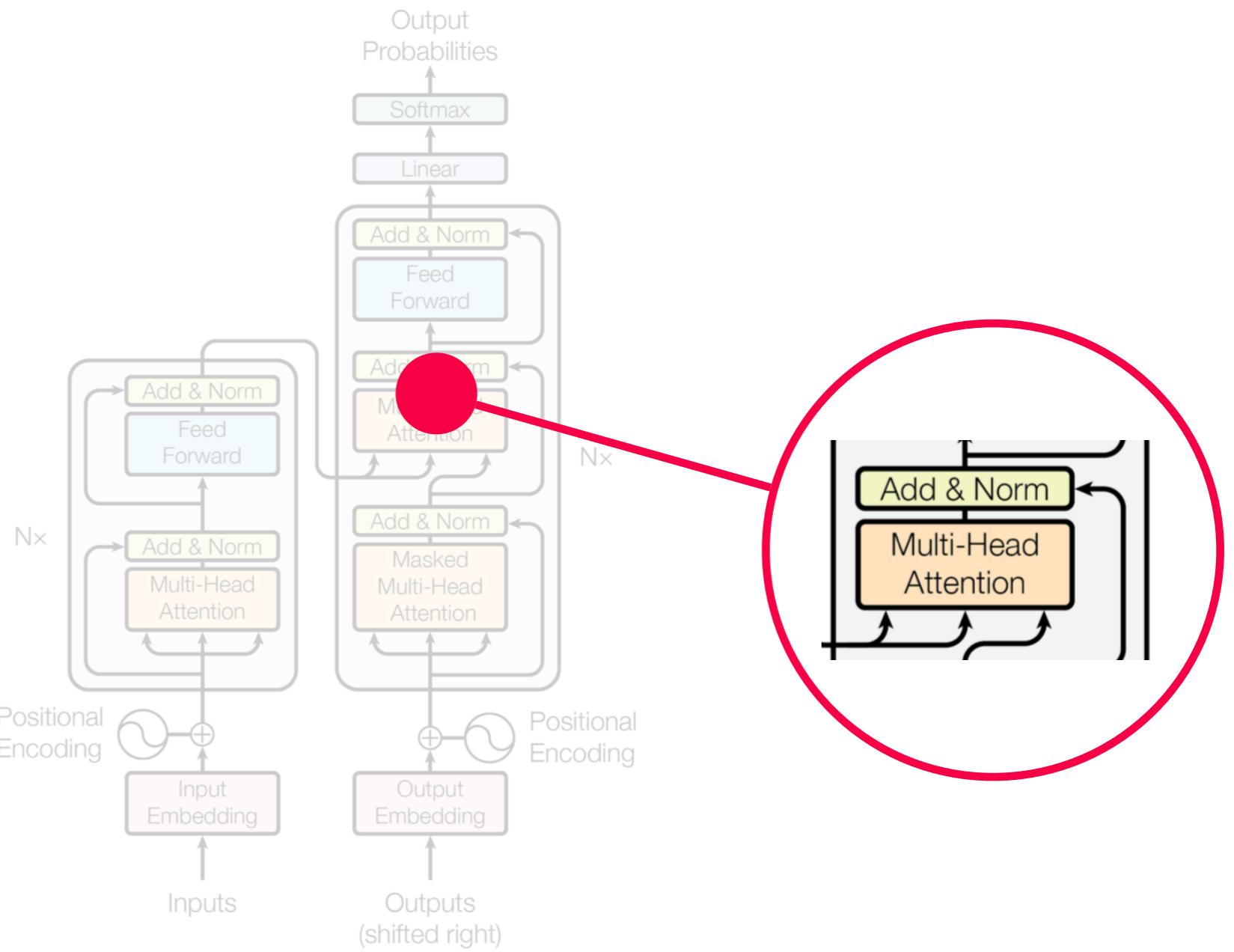


Figure 1: The Transformer - model architecture.

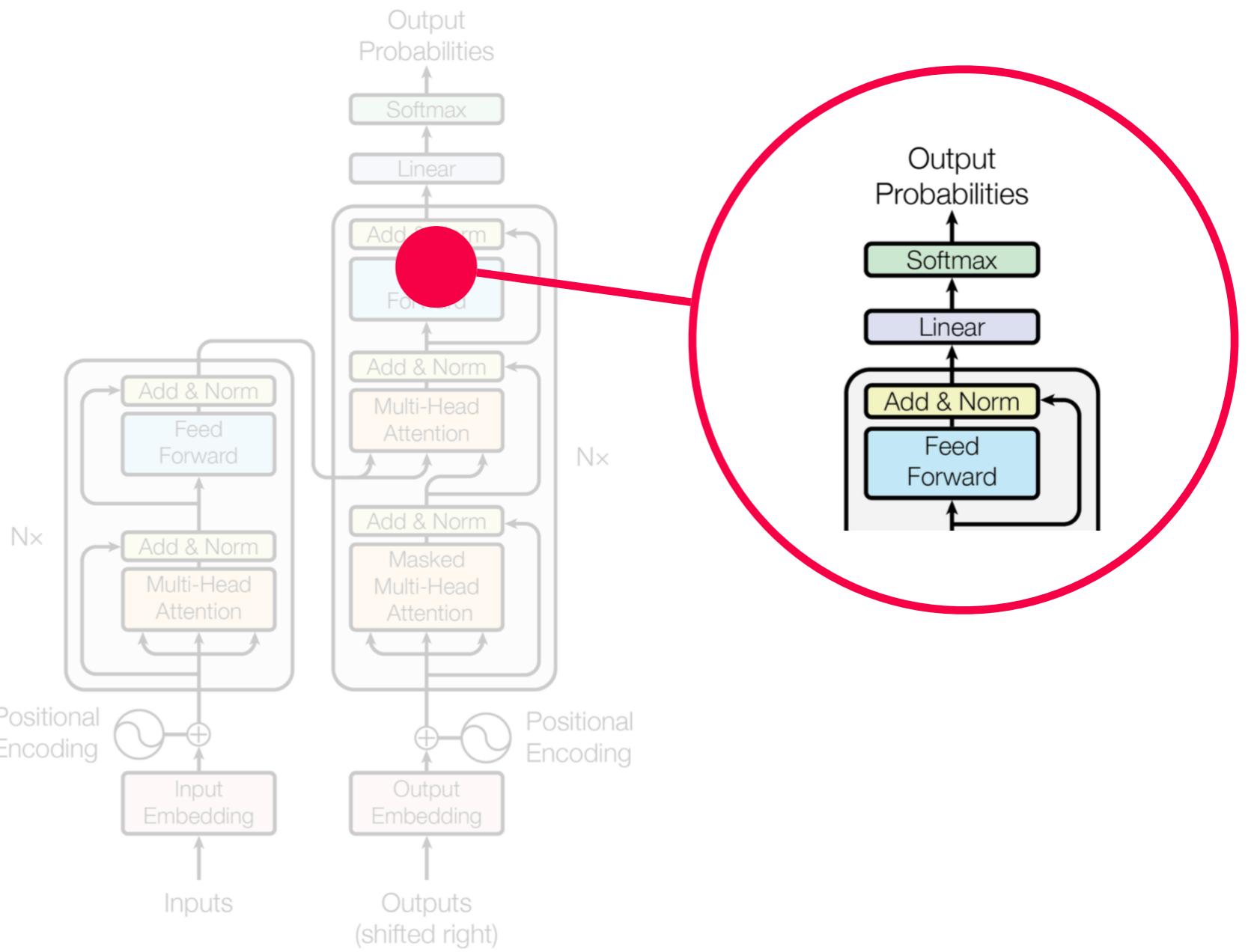


Figure 1: The Transformer - model architecture.

# ENCODER UND DECODER KÖNNEN UNABHÄNGIG VERWENDET WERDEN.

## Encoder-only models

dienen lediglich dazu numerische Repräsentationen des Textes zu erhalten; diese können dann zum Beispiel für eine Klassifizierung verwendet werden, **BERT**

## Decoder-only models

werden zur Textgenerierung verwendet

## Encoder-decoder models

auch sequence-to-sequence Modelle genannt, dienen der Übersetzung oder Textzusammenfassung

---

# ZUM SCHLUSS: **SERVICE HINWEISE**

---

# GÄNGIGE NLP PYTHON LIBRARIES

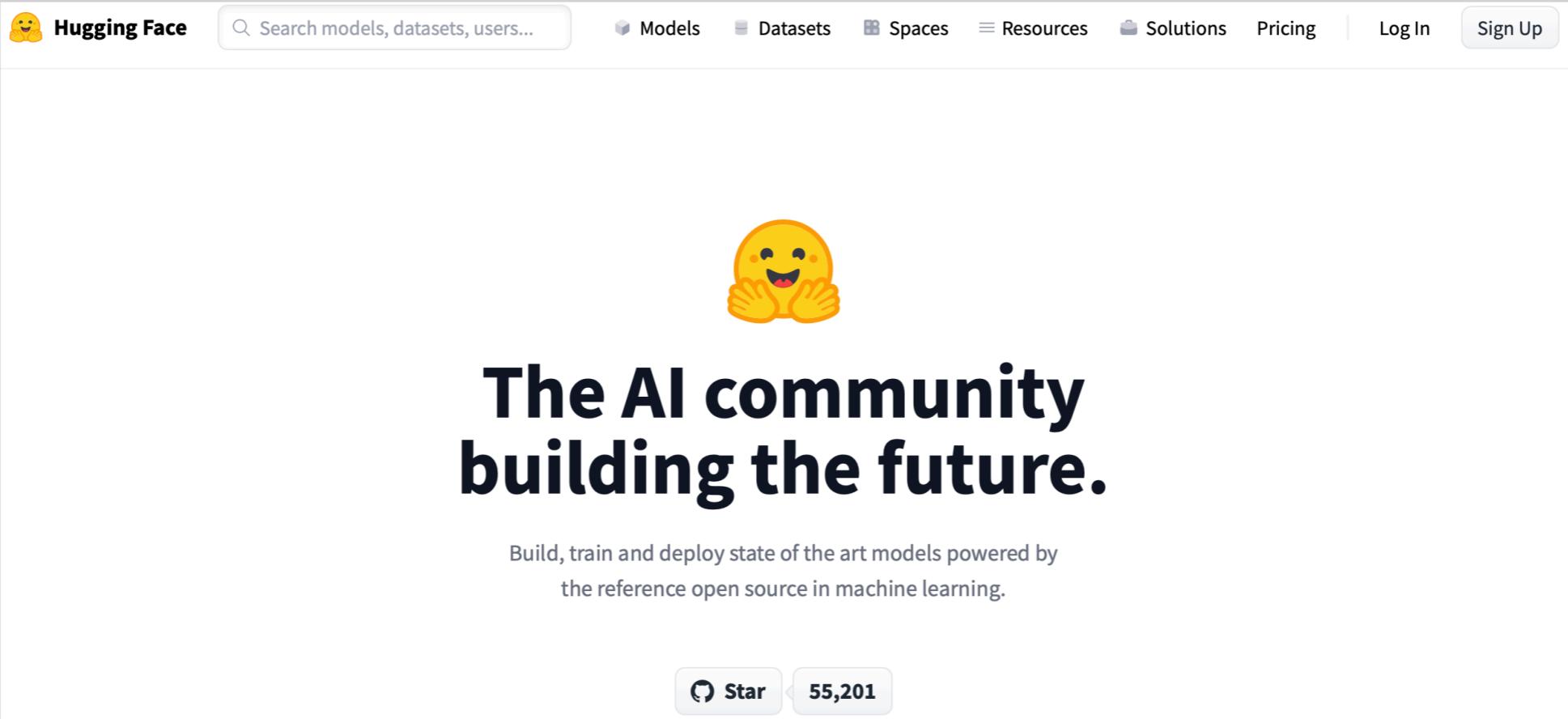
spaCy

NLTK

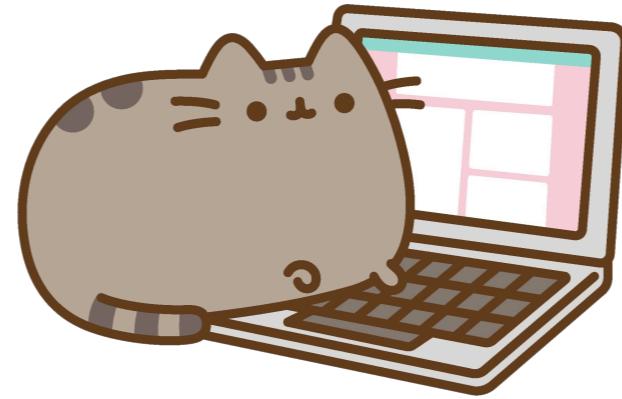
TextBlob

GenSim

# HUGGING FACE MACHT TRANSFORMER MODELLE FÜR DEN ANWENDER ZUGÄNGLICH!



The screenshot shows the homepage of the Hugging Face website. At the top, there is a navigation bar with a yellow smiley face icon labeled "Hugging Face", a search bar containing "Search models, datasets, users...", and links for "Models", "Datasets", "Spaces", "Resources", "Solutions", "Pricing", "Log In", and "Sign Up". The main content area features a large yellow smiley face icon with hands clasped together, followed by the text "The AI community building the future." in a large, bold, dark font. Below this, a subtitle reads "Build, train and deploy state of the art models powered by the reference open source in machine learning." At the bottom, there is a "Star" button with a count of "55,201".



---

# THANK YOU!

---

# RESSOURCEN

Das Hugging Face Tutorial:

<https://huggingface.co/course/chapter1/1>

Ein Podcast mit einem Interview mit dem neusten generativen Sprachmodell GPT-3:

<https://clearerthinkingpodcast.com/episode/073>