Chapter 10 | Sentiment Analysis

> "I can calculate the movement of the stars, but not the madness of men." ~ Sir Isaac Newton

10.1 The Field of Natural Language Processing

Natural Language Processing is one of the fastest-evolving fields in AI and machine learning. It might also be the shortest path to understand intelligence. When we think of an intelligent machine, we imagine a machine that has the language skills to communicate with us.

Alan Turing, in his famous 1950 paper on Computing Machinery and Intelligence ("I.—COMPUTING MACHINERY AND INTELLIGENCE | Mind | Oxford

Academic"), proposes to answer the question "Can Machine Think?" with an Imitation Game (now called the Turing test) based on language. A machine that can have a natural conversation with a human would be considered a thinking machine. Solving AI would, therefore, be equivalent to solving NLP.

Solving NLP involves many practical tasks that should be useful beyond looking for artificial general intelligence. In this chapter, we review some of these tasks and go over the different models used in modern deep learning NLP, including the GPT-3 model.

10.2 Language Tasks

NLP tasks are as diverse as the different uses of natural language. We present a non-exhaustive list of tasks: Question answering, machine translation, named entity extraction, coreference resolution, semantic role labeling, sentiment analysis, and textual entailment.

10.2.1 Question Answering

A question-answering task tests the reading comprehension of an NLP system. The NLP system should be able to answer questions. The prevalent benchmark is the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). It contains a list of 100,000 questions with answers identified as a segment of text (a span) in a Wikipedia entry. For instance, it answers "gravity" to the question "What causes precipitation to fall?".

The latest version, SQuaD, also includes 50,000 unanswerable questions. If the question does not have answers, a system should not offer one. An NLP system is given the question and retrieves the answer from the Wikipedia articles. It is evaluated according to its F1 score (F1 Score = 2*(Recall * Precision) / (Recall + Precision)).

10.2.2 Machine Translation

Machine translation is one of the most popular applications of NLP and is used in tools such as Google Translate or on Facebook to translate posts. Datasets used for machine translation are provided by the Workshop on Statistical Machine Translation (WMT) ("Translation Task - ACL 2014 Ninth Workshop on Statistical Machine Translation"). They include the WMT2014 English-German dataset and the WMT2014 English-French dataset.

The models are evaluated with the BLEU score, which considers human translation the benchmark. The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) is

a precision measure. It counts the matches of 4-grams to the human translation and makes adjustments for the translation length. The BLEU score is highly correlated with the human judgment of translation quality.

10.2.3 Named Entity Extraction

Named entity extraction identifies named entities in a text and assigns them to categories such as persons, organizations, locations, or miscellaneous entities. This task is helpful to search, reference, or classify documents. It has to identify the named entities, which can be one or several tokens, such as "the United States of America," and then classify the named entity correctly. It is evaluated according to its F1 score. A benchmark database is the Reuters RCV1 corpus ("Reuters Corpora @ NIST") with annotated entity classifications.

10.2.4 Coreference Resolution

Coreference resolution involves linking words referring to the same entity, especially pronouns, in a sentence. For instance, a benchmark database is the OntoNotes coreference annotations ("OntoNotes Release 5.0 - Linguistic Data Consortium"). It is evaluated according to its F1 score. An example of coreference resolution is the Winograd Schema Challenge. In the sentence, "The city councilmen refused the demonstrators a permit because they [feared/advocated] violence." Depending on which verb is used, "they" refers to either the "city councilmen" ("feared") or "demonstrators" ("advocated"). So, some deep understanding of the sentence seems to be required to identify the correct coreference. The Winograd Schema Challenge has been compared to the Turing Test.

10.2.5 Semantic Role Labeling

Semantic role labeling consists of labeling words according to their role around a predicate in a sentence. For instance, The Proposition Bank or PropBank ("The Proposition Bank (PropBank)"), built on top of the Penn Treebank ("Treebank-3 - Linguistic Data Consortium"), has a list of annotated sentences with verb predicates and defined roles for each argument of the predicate. The roles are specific for each verb predicate. For the predicate "agree," the roles are "Agreer," "Proposition," and "Other entity agreeing." Other familiar labeling sources are FrameNet, which focuses on frames and frame elements instead of a verb predicate; OntoNotes, which builds on top of the Penn Treebank for syntax; and Propbank for predicate-argument structure.

10.2.6 Sentiment Analysis

Sentiment analysis deals with the polarity, positive or negative, of a sentence or piece of text. It can be applied to movie reviews, product reviews, written reports, news articles, social media posts, or customer voice interactions. A standard database with annotated sentiments is the Stanford Sentiment Treebank ("Treebank-3 - Linguistic Data Consortium"), which uses around 11,000 sentences from movie reviews. Each movie review falls into one of five categories from very negative, negative, somewhat negative, neutral to moderately positive, positive, and very positive, as classified by Amazon Mechanical Turks. A Bag of words approach can be used where each word is given a sentiment score but is sometimes insufficient because it lacks context and order.

10.2.7 Textual entailment

Textual entailment is the relationship between a text and a hypothesis. From a text or a fact, an NLP system has to evaluate if a hypothesis is True (entailment), False (contradiction), or Neutral. A benchmark is the Stanford Natural Language Inference (SNLI) ("The Stanford Natural Language Processing Group"). It has 570,000 entries of text, judgments (entailment, contradiction, or neutral), and hypotheses. For instance, the text could be "A soccer game with multiple males playing." The hypothesis is "A soccer game with multiple males playing." The judgment is "entailment" because the text backs the hypothesis. If the text is "A black race car starts up in front of a crowd of people." and the hypothesis is "A man is driving down a lonely road." then the judgment is "contradiction."

10.2.8 Other Tasks

There are many other NLP tasks such as speech recognition (used by personal assistants such as Siri or Alexa), text-to-speech to read texts, text summarization to summarize news articles, reports, or books, text classification to screen for email spam, offensive contents, or identify authorship, information extraction to collect data from web pages or online documents, information retrieval to find relevant documents or pieces of information (used in search engines in Google, YouTube or Amazon).

10.3 Classical NLP Modelling

10.3.1 Symbolic NLP

Teaching a computer vocabulary, syntax, and grammar to solve language tasks is possible. This approach is symbolic NLP and uses parsing techniques to identify the words, their roles, and their meanings (Part-of-Speech or POS tagging). Because of the complexity and ambiguity of language and its relative free form, it is difficult to make a hand-written inventory of all the rules required to understand and generate some language.

Another approach is to learn language probabilistically, using a statistical language model trained on real-world data. The statistical approach has gained the upper hand because of the considerable amount of digital text available with corpora of millions, billions, and even trillions of words and the large availability of computing power. At the same time, the symbolic paradigm has not made meaningful progress in real-world applications. Despite its success, MIT Professor Noam Chomsky has been very critical of the statistical approach. He was quoted as saying:

"It's true there's been a lot of work on applying statistical models to various linguistic problems. I think there have been some successes, but a lot of failures. There is a notion of success, which I think is novel in the history of science. It interprets success as approximating unanalyzed data." ("Pinker/Chomsky Q&A from MIT150 Panel")

Norvig ("On Chomsky and the Two Cultures of Statistical Learning") has an interesting article addressing his criticism. In particular, he points out the empirical success of these models applied to search engines (Norvig works at Google), speech recognition, machine translation, and question-answering.

10.3.2 Language Model

A language model describes the probability distribution of words. It is a statistical representation of language. It answers the question of the probability that a word appears after a sequence of words or the probability that a sentence was said vs. another. This is a compelling approach to developing language applications because it can leverage existing textual data and can tell, for instance, if a sentence is grammatically correct or logical because correct and logical sentences are more likely to occur in the data.

10 3.2.1 Bag of Words

The simplest language model is the bag-of-words model, where only the frequency of each word matters, neither the ordering nor the presence of other words. It is a poor model to generate sentences, but it is helpful to measure sentiment or classify text. Suppose some words tend to appear more frequently in a negative sentence. In that case, their presence can indicate that a sentence is likely negative, using the Bayes formula of conditional probabilities.

10.3.2.2 N-gram Models

A more advanced approach than the bag-of-words is the N-gram model. In the N-gram model, the probability of each word is conditional (depends) on the previous N-1 words. A bigram model accounts only for the previous word; a 3-gram model will account for the last two words, etc. Given these conditional probabilities, the probability of a full sentence can be calculated thanks to the law of iterated expectations. It will be expressed as a simple product of conditional probabilities or as the sum of logarithmic probabilities if logarithms are used. N-gram models can be used for spam detection, sentiment analysis, or document classification.

10.4 Deep NLP Modelling

10.4.1 Word Embedding and Word Vectors

The previous language models do not compare words. Two similar or related words should be close in some dimension, and word vectors allow these comparisons. Word vectors are also called word embeddings. Two successful approaches have been Glove and Word2Vec.

10.4.1.1 Glove (Global vectors)

Glove (Pennington et al., 2014) was developed at Stanford to construct vector representations of words. It is based on the co-occurrence of words. Co-occurrence means that words occur together in the same sentence. An unsupervised machine learning model on a corpus to estimate the co-occurrence of pairs of words. The word vectors are estimated so that the dot product of two-word vectors is equal to the probability of co-occurrence. Thanks to this vector representation, relationships between words can be seen, such as man to woman and king to queen.

10.4.1.2 Word2Vec

Word2Vec (Mikolov et al., 2013) was developed at Google and also aims to create word vectors where similar words have close representations. Closeness is measured with the continuous bag of words (CBOW) or the continuous skip-gram. With a continuous bag of words, we predict a word according to its context. With continuous skip-grams, a word predicts the context words surrounding it. We use a two-layer neural network to estimate each model.

10.4.2 The GLUE Benchmark

The General Language Understanding Evaluation (GLUE (Wang et al., 2019)) benchmark is a set of tests to evaluate NLP models on different tasks of sentence understanding. Some tasks are based on individual sentences, some others on pairs of sentences.

Task	Corpus	Metrics	Domain
Acceptability (grammatical English)	CoLA Corpus of Linguistic Acceptability	Matthew's correlation	Miscellaneous
Sentiment	SST-2 The Stanford Sentiment Treebank	Accuracy	Movie reviews
Paraphrase	MRPC The Microsoft Research Paraphrase Corpus	Accuracy/F 1	News
Sentence similarity	QQP The Quora Question Pairs	Accuracy/F 1	Social QA questions
Paraphrase	STS-B The Semantic Textual Similarity Benchmark	Pearson/ Spearman correlation	Miscellaneous
Natural language inference	MNLI The Multi-Genre Natural Language Inference Corpus	Matched accuracy/ Mismatche d accuracy	Miscellaneous
Question answering/Natural language inference	QNLI The Stanford Question Answering Dataset	Accuracy	Wikipedia

Natural language inference	RTE The Recognizing Textual Entailment	Accuracy	News, Wikipedia
Coreference/Natural language inference	WNLI The Winograd Schema Challenge	Accuracy	Fiction books

Table 10.1. The GLUE tasks.

Because the new generations of NLP models tend to have superhuman performance in some tasks from the GLUE benchmark, Wang and his colleagues (Wang et al., 2020) have introduced SuperGLUE with more difficult and more varied tasks and human benchmarks.

10.4.3 Recurrent Neural Network (RNN)

RNNs (Elman, 1990) are a type of neural network that allows efficient modeling of sequences, such as time series or text data.

In a basic RNN, for each step t, an input vector x_t is combined with a hidden vector or layer h_{t-1} to produce an updated vector h_t , generating the vector y_t. In the next step $t+1$, the new input vector $x_{t+1}$is combined with h_t from the previous step to produce the new output vector y_{t+1}. The relationship between $x_{t+n}, h_{t+n-1}, h_{t+n}$, and x_{t+n} is independent of n which makes it more efficient with fewer parameters to estimate.

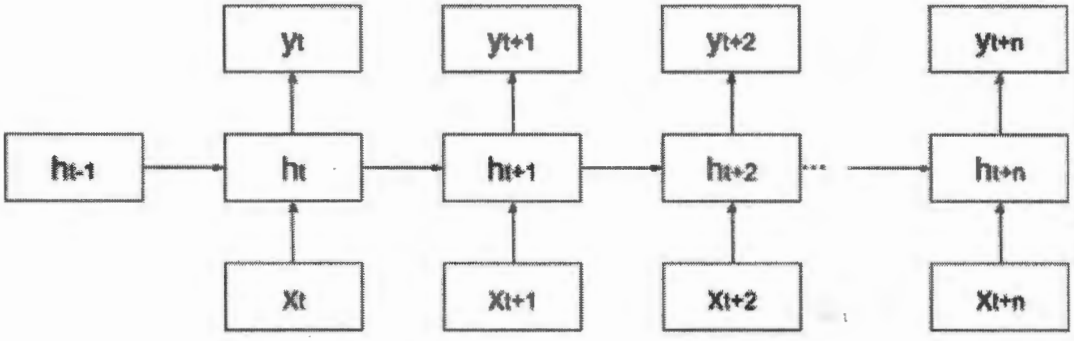

Figure 10.1. Recurrent Neural Networks

The hidden layer h_{t+n} keeps the memory of the previous step layers $h_{t+n-1}, h_{t+n-2}, \ldots, h_o$. The parameters are estimated by back-propagation through time, starting from the last period and moving back to the initial values of each layer.

As for a neural network with too many layers, the RNN can suffer from vanishing gradients (the gradient becoming smaller and smaller as we go back in time) or exploding gradients (the gradient becoming larger and larger). The LSTM model has been created to address this issue.

10.4.4 Long Short-Term Memory (LSTM)

LSTM was introduced by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber, 1997). The LSTM uses a carry or memory cell c_t, which depends on an input gate i_t and a forget gate f_t. The output depends on an output gate o_t.

The memory cell carries information from one step to the next but is more flexible than the hidden state. The information is copied with some adjustments. The memory cell depends on the following:

- an input gate i_t: the input gate modulates the information from the input layer x_t and the hidden layer h_t
- a forget gate f_t: the forget date can erase some memory cell information

Therefore, the memory cell can forget some memory with the forget gate and use some new memory content thanks to the input gate. σ is the sigmoid function.

$$c_{t+1} = f_t \odot c_t + i_t \odot \sigma(b \quad + U \quad x_t + W \quad h_t)$$

$\odot$ is the element-wise multiplication.

The output uses an output gate o_t. The output gate modulates the memory cell c_t to transform it into an output vector y_t. The output is calculated as:

$$y_t = o_t \odot tanh(c_t)$$

The input gate, the output gate, and the forget gate are updated with a sigmoid function σ:

$$i_t = \sigma(b_i + U_i x_t + W_i h_{t-1})$$

$$o_t = \sigma(b_o + U_o x_t + W_o h_{t-1})$$

$$f_t = \sigma(b_f + U_f x_t + W_f h_{t-1})$$

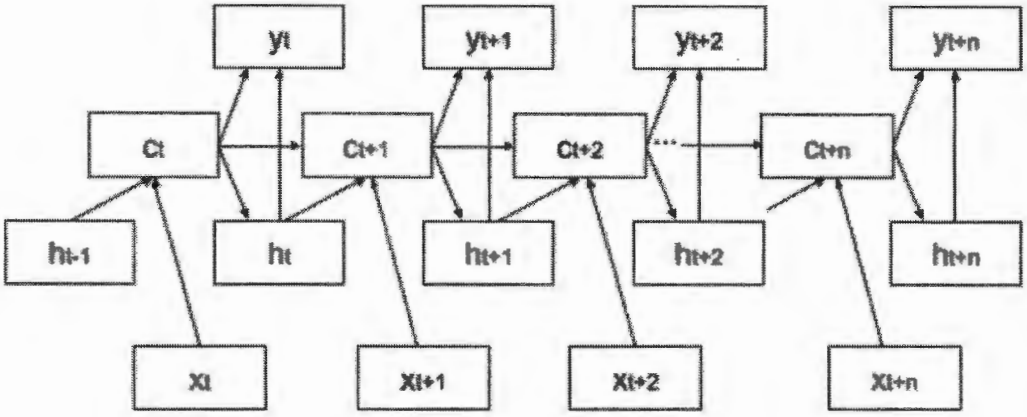

Figure 10.2. LTSM

The output y_t will depend on the hidden state h_t and memory cell c_t.

Compared to the simple RNN, the input layer x_t does not feed the hidden layer h_t directly, but indirectly through the memory cell c_t. The hidden layer h_{t-1} does not feed into the next hidden layer h_t directly but only indirectly through the memory cell c_t.

10.4.5 Bi-directional LSTM

With the bidirectional LSTM (Graves and Schmidhuber, 2005), the same sequence is analyzed in reverse, and two LSTM outputs are combined by concatenation, sum, or product (Figure 10.3).

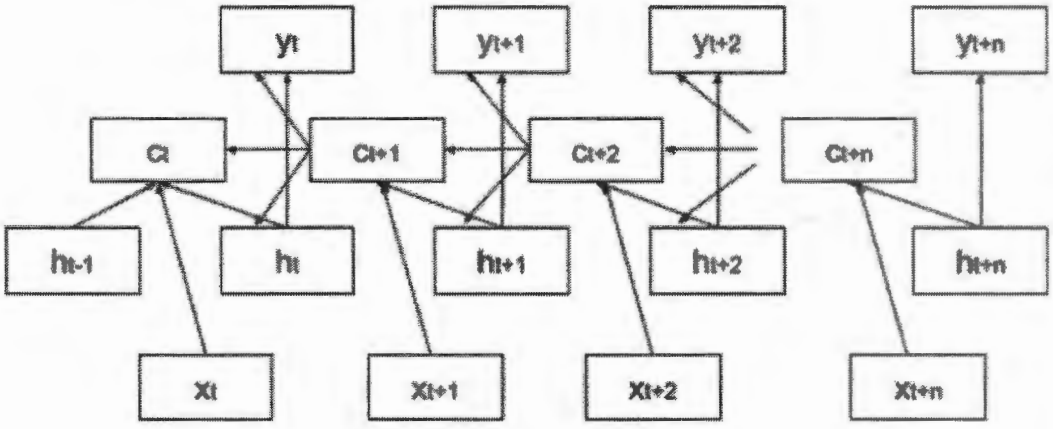

Figure 10.3. Bidirectional LSTM

10.4.6 Gated Recurrent Units (GRU)

The GRU was introduced by Cho (Cho et al., 2014) to simplify the LSTM. There is no more hidden layer. The output layer depends on an update gate u_t and a reset gate r_t.

The update gate u_t and the reset gate r_t are updated with a sigmoid function σ:

$$u_t = \sigma(b_u + U_u x_t + W_u y_t)$$
$$r_t = \sigma(b_r + U_r x_t + W_r y_t)$$

The output layer y_t is then updated as follows:

$$y_{t+1} = u_t \odot y_t + (1 - u_t) \odot \sigma(b \quad + U \quad x_t + W r_t \odot \quad y_t)$$

Updating and resetting to a new value is determined in a single equation and bypasses a memory cell and a hidden layer.

10.4.7 ELMo (Embeddings from Language Models)

In traditional word embedding, a word can have only one meaning. ELMo, proposed in 2018 by the Allen AI Institute and the University of Washington (Peters et al., 2018), improves on traditional static word embeddings such as GloVe by using the context of word usage. It constructs a vector representation of words based on the parameters of a bidirectional LTSM model trained on a large text corpus. The representation depends on the whole sentence in which the word appears. These are contextualized representations since they rely on the context of the word.

The parameters are from all the layers of the biLSTMs and not only from the last layer. The parameters from the upper layers help understand the context, while the parameters from the lower layers help to understand the syntax.

ELMo can be integrated to improve NLP tasks. The biLSTM model is run on the text, and the ELMo representations and the status word representations are both fed into the supervised NLP tasks. ELMo improves the performance of many tasks, such as question answering, text entailment, semantic role labeling, or coreference resolution.

10.4.8 Attention Model

10.4.8.1 Attention

The concept of attention allows the dynamic association of a word or token in a sequence to some words or tokens in another or the same sequence. This allows richer associations that do not depend on specific locations of the target words; in particular, it can relate a word to words that are not in close proximity. This is useful in translation, for instance, where a meaningful word can be at the beginning of a sentence and still be helpful to translate a word appearing at the end.

Attention (Vaswani et al., 2017) uses the concept of Queries, Keys, and Values. A Query is what we are looking for, the Key gives the location of what we are looking for, and the Value is the query's result. A Query is, for instance, a word; the Key is a page in a dictionary where the word appears, and the Value is the translation of the word.

A word represented as an embedding vector x is multiplied (matrix dot product) with a query weight matrix W^Q to produce Queries Q, with a key weight matrix W^K to produce Keys K. These matrices Q, K, V are then combined and transformed into probabilities (through a softmax function and after normalization) to emphasize attention to specific values or tokens in the same sequence. The values V are calculated as the dot product between the initial token and a value weight matrix W^V.

The self-attention vector is then:

$$attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$$

d_k is the dimension of the key vectors, and its square root is a normalization factor.

We can represent it in a picture:

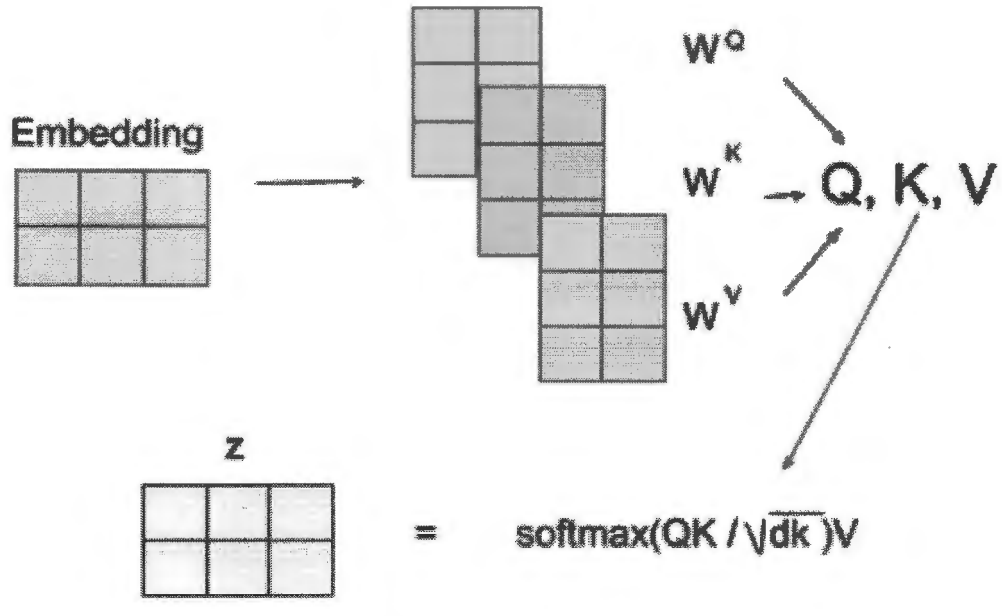

Figure 10.5. Attention mechanism

$attention(Q, K, V)$ is called a dot product attention (here, a scaled dot product attention because of the scale factor). It is called self-attention if the target sequence is the same. It is causal attention if attention cannot look forward, and a mask is used to eliminate any forward-looking attention. It is bimodal attention if attention can look backward and forward (two directions).

To have causal attention, we just add to QK^T a triangular matrix M with 0 everywhere and $-\infty$ in the upper triangular area of the matrix.

10.4.8.2 Multi-Head Attention

The procedure can be repeated several times with different $W^Q{}_i$, $W^K{}_i$, $W^V{}_i$ matrices to create several self-attention vectors or "multi-head" attention. These vectors are then concatenated and multiplied by another weight matrix W^O to produce a single self-attention vector.

$$z_i = attention(Q_i, K_i, V_i) \text{ for } i = 1, .., h \text{ if there are } h \text{ heads.}$$

Then

$$Multihead(Q, K, V) = Concat(z_1, \ldots, z_h)W^O$$

10.4.8.3 Transformers

Transformers (Vaswani et al., 2017) use multi-head attention, add and normalize layers and feed-forward layers. A Transformer layer can combine an encoder and a decoder or use only an encoder (as in BERT models) or only a decoder (as in GPT models).

They use word embeddings inputs. Word embeddings are vector representations of words. A vector of fixed dimensions represents each word. The word embedding is a list of such vectors. The list size is usually set to be equal to the longest sentence in a text.

10.4.8.3.1 Encoder

The encoder has four layers. As input, it uses an embedding added to a positional encoding. The positional encoding indicates the location of each word vector. In the encoder, the input goes through a multi-head attention layer, which encodes each word vector with other vectors it needs to pay attention to. It is then added and normalized with the original layer input to preserve some memory of the input. It goes to a feed-forward layer and again another "add & normalize" layer. The output is then fed into a multi-head attention layer of the decoder.

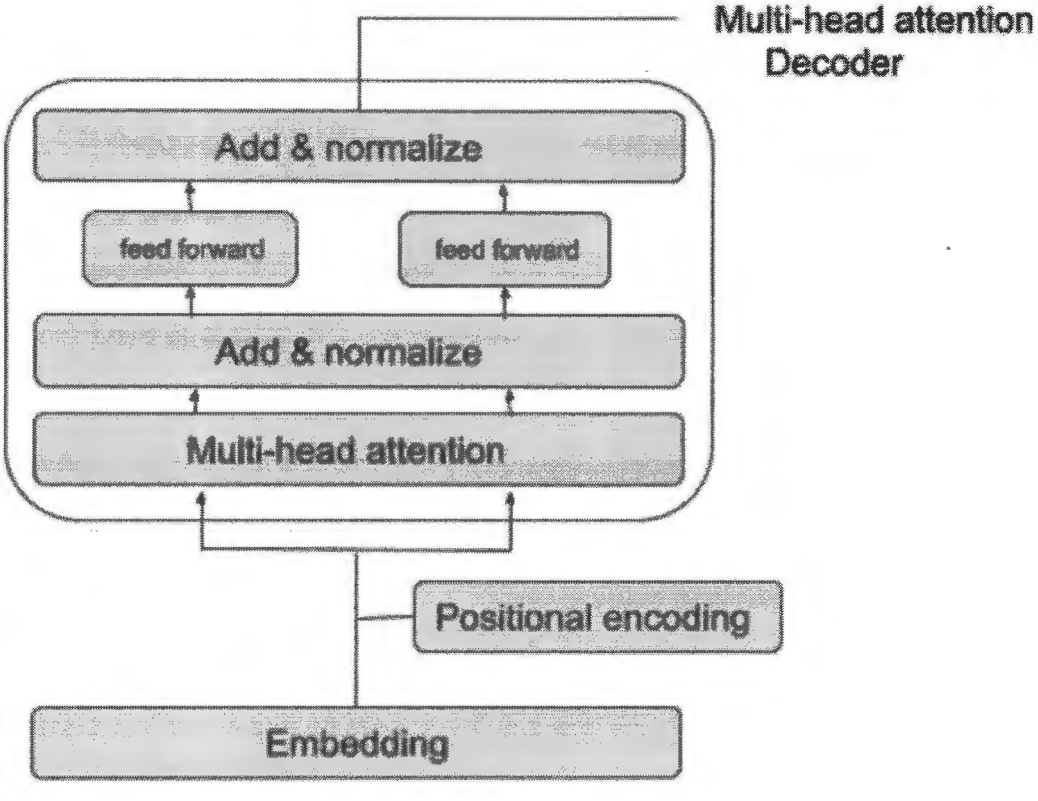

Figure 10.6. Transformer Encoder Layer

10.4.8.3.2 Decoder

The decoder is very similar to the encoder, but it uses a masked multi-head attention layer to pay attention only to past word vectors.

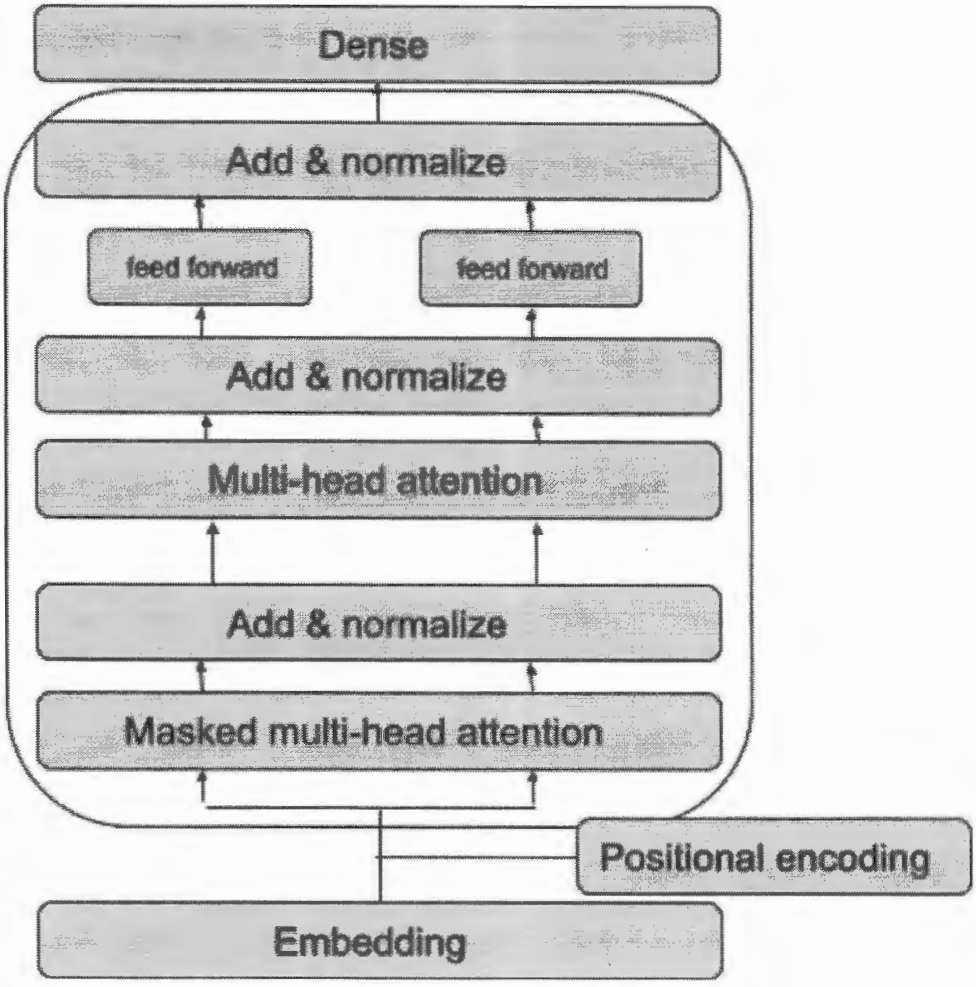

Figure 10.7. Transformer Decoder Layer

10.4.8.3.3 Transformer

The Transformer layer uses a word embedding as input to the encoder. The result is fed into the decoder along with the output word embedding. The output word embedding is inputted into the first decoder layer repeatedly as new words get outputted.

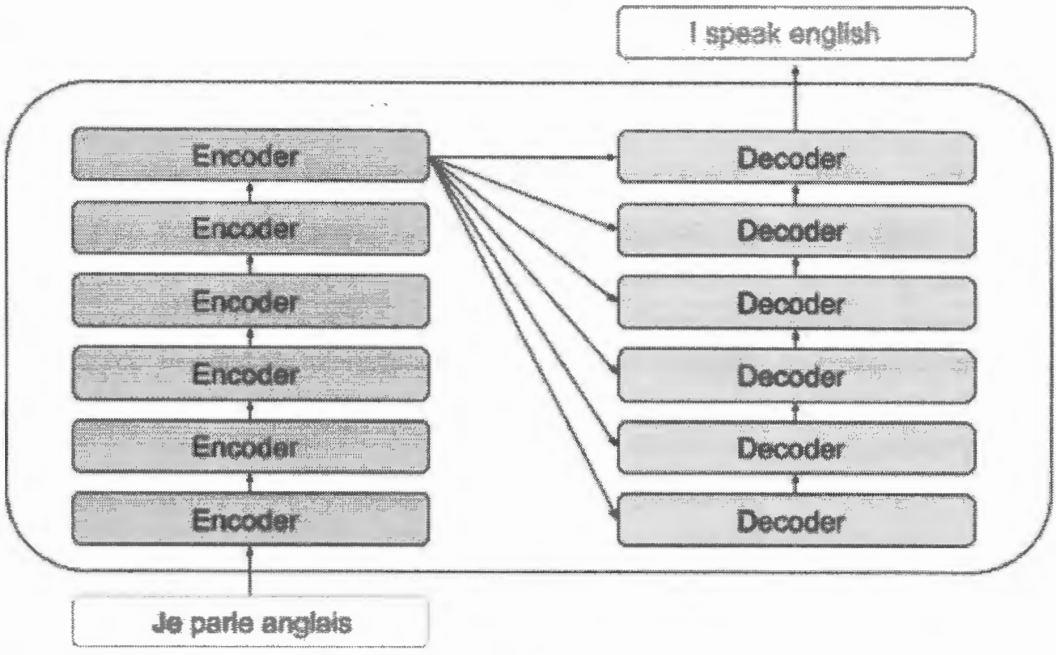

Figure 10.8. Transformer

Transformers have launched a new wave of pre-trained language models such as BERT and GPT-3. We review some of them.

10.4.8.4 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained language model introduced by Google in 2018 (Devlin et al., 2019) that can be fine-tuned to perform many everyday NLP tasks, such as the ones from the BLUE benchmark. Contrary to ELMo, which uses the new embeddings as new features, BERT requires minimal re-training.

BERT uses transformers with layers of decoders. It is trained first to identify randomly masked words (Masked Language Model) in a sentence using their contexts, words from the left and the right of the mask, and then to predict the following sentence (Next Sentence Prediction). It is, therefore, bidirectional, contrary to GPT style models, which are unidirectional. BERT uses a multi-layer bidirectional Transformer encoder.

There are two versions of BERT: BERT base and BERT large. BERT base has 12 layers of size 768, 12 self-attention heads, and 110M parameters. BERT large has 24 layers of size 1024, 16 self-attention heads, and 340M parameters. BERT is described in Figure 10.9.

BERT is pre-trained on a Corpus made of the BooksCorpus (800M words) and the English Wikipedia (2,500M words), representing a total of 3.3 billion words. The text

goes through WordPiece tokenization and then runs through a masking step where tokens are masked at the rate of 15%. The token is replaced by [MASK] 80% of the time, by a random token 10% of the time, and remains the same 10% of the time. [MASK] is not used 100% of the time because it does appear in the fine-tuning step. BERT then goes through the Next Sentence Prediction step, in which pairs of sentences can either be paired correctly with the label [IsNext] 50% of the time or with the label [NotNext].

BERT is then fine-tuned on specific tasks. Most of the hyperparameters remain the same. The model parameters are re-estimated. The input can be pairs of sentences in the case of machine translation or question answering, and the output will be some token representations to be fed into a single additional task-specific layer.

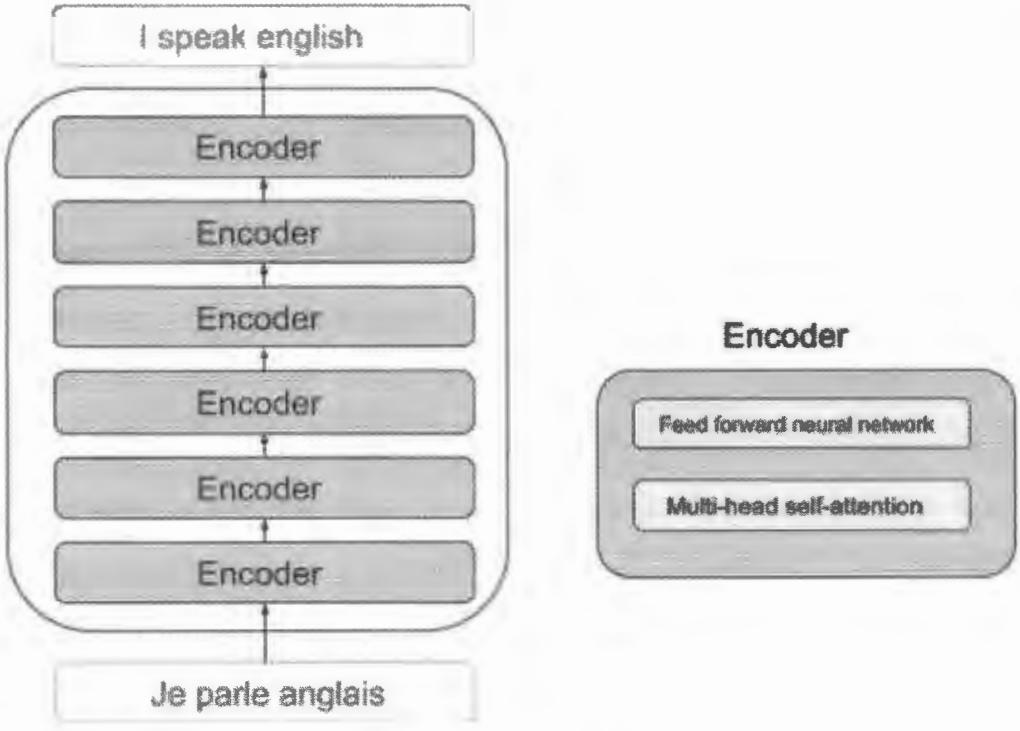

Figure 10.9. BERT

10.4.8.5 RoBERTa

RoBERTa (Robustly optimized BERT pretraining approach (Liu et al., 2019)) is a reimplementation of BERT by Facebook where they change the following: a more extended training period, bigger batches, more data, no next sentence prediction, longer sequences, and dynamic masking pattern on the training data. The modifications

are significantly improving the model performance, and it achieves state-of-the-art results on GLUE, RACE, and SQuAD.

10.4.8.6 XLNet

XLNet (Dai et al., 2019) is an improvement of the BERT model from Google and Stanford University. It uses a Transformer architecture but uses an auto-regressive approach without masking. It performs token permutations to feed into the encoder layer and tries to predict each token. XLNet also includes ideas from Transformer-XL, such as the relative positional encoding scheme and the segment recurrence mechanism into pretraining. XLNet performs better than BERT on many NLP tasks, including question answering, natural language inference, sentiment analysis, and document ranking.

10.4.8.7 ELECTRA

The ELECTRA (Clark et al., 2020) model proposes to use an alternative to masking, which is more sample-efficient than in BERT. It replaces tokens randomly with alternatives generated by a neural network, and the task is to detect these replacements. ELECTRA outperforms BERT on the GLUE benchmark when both run with the same model size, data, and computation. It also outperforms XLNet and ROBERTa with the same amount of computing.

10.4.8.8 T5

T5 (Raffel et al., 2020) is a unified framework for language modeling based on the original transformer architecture with very few changes. It is framed as a text-to-text problem. They use a new cleaned-up data set, the "Colossal Clean Crawled Corpus"' and achieve state-of-the-art results on many NLP benchmark tasks such as summarization and question-answering text classification. The T5 needs to be fine-tuned by changing all the pre-trained weights.

10.4.8.9 GPT-3 (Generative Pre-Training)

GPT-3 (Brown et al., 2020) is a language model that can be used for many downstream tasks, such as question answering, text completion, text generation, and neural machine translation. GPT-3 is the third generation of the Generative Pre-Training (GPT) model (Radford et al., 2018). GPT-3 is described in Figure 10.10 below.

The original GPT model is pre-trained on a large corpus of text using unsupervised learning and transformers. Each layer of the GPT model is a transformer decoder

layer. A decoder layer contains an attention layer and a feed-forward neural network. The attention layer is a self-attention layer. The masked attention layer is a masked multi-head self-attention layer that cannot look forward. The model is then fine-tuned to specific tasks with supervised learning. The GPT-3 model skips the fine-tuning step.

GPT-3 has 175 billion trainable parameters, 96 layers, 12,288 units in each bottleneck layer, and 96 attention heads with 128 units each. Performance increases with the number of parameters. Because of its size, GPT-3 can perform well without fine-tuning. The weights do not need to be re-estimated for a new task.

GPT-3 is trained on a combination of five data sets: filtered Common Crowl (410 billion tokens), WebText2 (19 billion), Books1 (12 billion), Books2 (55 billion), and Wikipedia (3 billion). Some datasets are seen several times if they are of higher quality.

GPT-3 is evaluated with few-shot learning, one-shot learning, and zero-shot learning. An X-shot learning means the model is given X examples before returning an answer to a query. GPT-3 improves the state-of-the-art results on several benchmark tasks, such as sentence completion, question answering, and machine translation to English, but still falls short on some others, such as common-sense reasoning and reading comprehension. For benchmarks such as SuperGlue, it falls short of the best-fine-tuned models. GPT-3 shines at news article generation. Humans were only 52% accurate at guessing that an article was written by GPT-3 instead of a human.

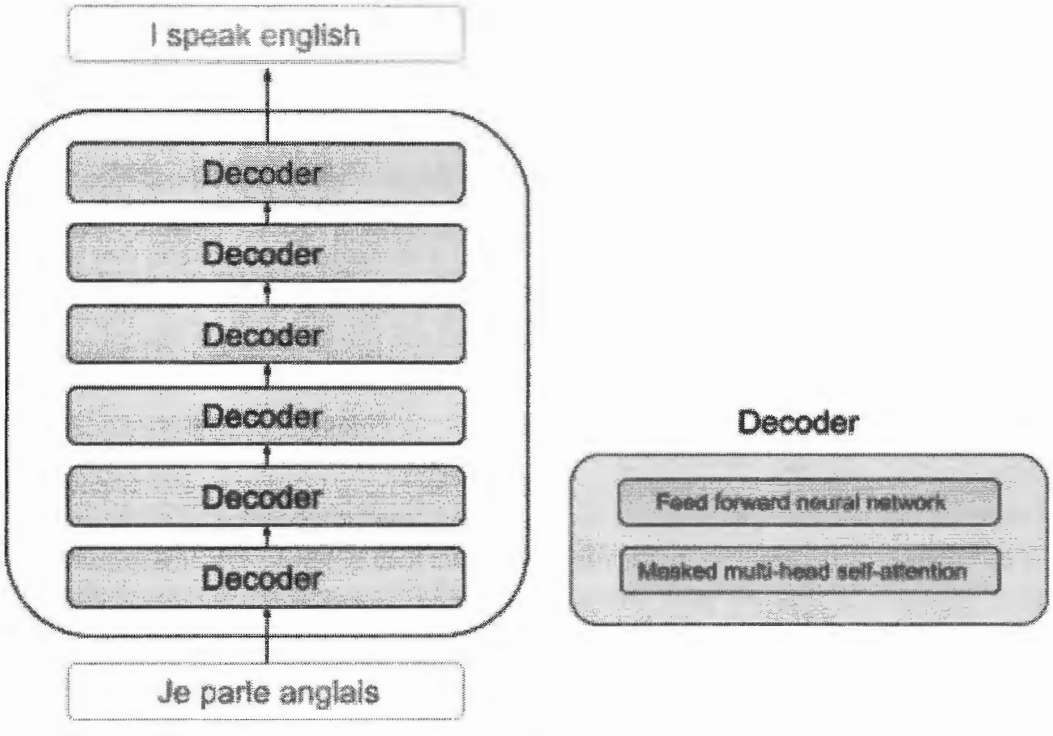

Figure 10.10. GPT-3

After GPT-3 comes GPT-3.5, ChatGPT, and GPT-4 (OpenAI, 2023) . We review more models on Generative AI in Chapter 13.

10.4.8.10 Megatron-Turing NLG 530B

Megatron-Turing NLG (Smith et al., 2022) is a GPT3-style model with 105 layers and 530 billion parameters developed by Microsoft and Nvidia. It was, as of October 2021, the most prominent language model in existence. Thanks to the parallelization of data, pipelines, and tensor-slicing (to mask elements in a tensor), it achieved state-of-the-art results in completion prediction, reading comprehension, commonsense reasoning, natural language inferences, and word sense disambiguation.

10.5 NLP and Sentiment Analysis

10.5.1 Sentiment Analysis

Sentiment analysis measured by text is a subfield of Natural Language Processing (NLP), a subfield of AI and machine learning. Applied to investing, AI can relate sentiment indicators to investment performance. A working hypothesis is that positive (negative) sentiment in news and financial reports should be associated with future relative positive (negative) stock returns as long as the sentiment is measured correctly and adapted to the financial domain. Another hypothesis is that extreme positive sentiment could lead to reversal, either on the downside when investors are too optimistic or on the upside when they are too pessimistic.

AI-based sentiment analysis systems need to understand the context of some opinions and be robust to the way an opinion is expressed. Ideally, these systems should work across different modalities (text, voice, image, and video) and remain consistent. A CEO who publishes a positive annual report and delivers a positive earnings outlook in a conference call with Wall Street analysts should lead to the same positive sentiment measure.

Sentiment analysis can also be more detailed beyond just being positive or negative. For instance, an AI should be able to tell that a firm is more positive about its prospects in one region compared to another one (e.g., United States vs. Europe or Asia) or in one market segment relative to another one (e.g., electric vehicles vs. internal combustion engine vehicles).

Sentiment analysis can also be applied to alternative data (Chapter 12), such as originating from social media (X formerly Twitter, Facebook, LinkedIn), blogs, online forums, and discussion boards such as Reddit, and product reviews published in news media or generated by individual consumers. Customer reviews and feedback, for instance, can be seen on e-commerce websites such as Amazon and YouTube videos published by private individuals or social media influencers.

In the context of finance, (Loughran and McDonald, 2011) explore the use of textual analysis to measure the tone of corporate 10-K reports in finance and accounting research. They examine the effectiveness of word lists, mainly the Harvard Dictionary, in classifying words as negative in financial contexts. They find that a significant portion of words identified as negative by the Harvard Dictionary are not considered negative in financial contexts, leading to misclassifications and noise in tone measurement. They create an alternative list of negative financial words and five other word lists (positive,

uncertainty, litigious, strong modal, and weak modal words) that better reflect the tone of the financial text. They link these word lists to various financial variables such as returns, trading volume, volatility, fraud, material weakness, and unexpected earnings and show that the word lists reveal significant relations with various financial variables and are explored in different contexts beyond 10-K reports.

(Mishev et al., 2020) show that contextual embeddings provided by NLP transformers outperform lexicons and fixed encoders, even when working with limited datasets. Additionally, they find that distilled versions of NLP transformers are suitable for production environments, offering comparable results to their larger teacher models.

10.5.2 Market-Level Sentiment Analysis

In a very influential article, (Baker and Wurgler, 2006) study investor sentiment and the cross-section of stock returns and provide some evidence of the effect of sentiment, especially on newer, smaller, unprofitable, high-growth, and more volatile stocks with data from the CRSP-Compustat database covering 1962 to 2001. They identify periods of high sentiment in the 60s, 70s, 80s, and the dot com boom in the late 90s and track the performance of stocks with characteristics related to firm size and age, profitability, dividends, asset tangibility, and growth opportunities and distress. The most difficult stocks to value are also the ones that are most sensitive to sentiment.

They measure aggregate sentiment with a composite index that covers the closed-end fund discount (difference between the Net Asset Values of the assets in the fund and its market price), NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium (log difference of the average market-to-book ratios of dividend payers and non-payers). The indicators are orthogonalized to the business cycle (regressed by business cycle indicators are replaced by their residuals) so that the sentiment index captures pure sentiment only.

They examine the effect of the aggregate sentiment index on the subsequent annual stock returns using sorting and regression analysis. They find that sentiment is high, stocks that are hard to value and unattractive to arbitrageurs (younger stocks, small stocks, unprofitable stocks, non-dividend paying stocks, high volatility stocks, extreme growth stocks, and distressed stocks) tend to underperform. This pattern attenuates or reverses when sentiment is low. The overvaluation effect during high sentiment periods is more pronounced at the extreme of the cross-sectional distribution.

How about textual information? (Tetlock, 2007) examines the relationship between financial news content, particularly in the Wall Street Journal's "Abreast of the Market" column, and daily stock market activity. He uses the General Inquirer, a computer-

assisted approach for content analyses of textual data, to analyze the content and identifies a pessimism factor in media reports, which is strongly related to negative words. He finds evidence that the media's pessimism can predict movements in stock market indicators. High levels of media pessimism are associated with downward pressure on market prices, followed by a return to fundamentals. Unusually high or low values of media pessimism forecast high market trading volume. Low market returns lead to high media pessimism. These findings suggest that media content serves as a proxy for investor sentiment and non-informational trading.

(Jiang et al., 2019) explore the role of manager sentiment, but based on the aggregated textual tone of corporate financial disclosures, in predicting future stock market returns. They construct a manager sentiment index and find that manager sentiment is a strong negative predictor of future stock market returns. Specifically, the manager sentiment index is negatively correlated with future stock returns, and a one-standard-deviation increase in manager sentiment is associated with a decrease in expected excess market returns for the next month. This predictive power remains robust even when compared to other macroeconomic predictors.

Instead of corporate financial disclosures, (Thorsrud, 2020) constructs a daily business cycle index that provides timely and informative insights into economic conditions using economic news data from Norway from 1988 to 2014. The author combines traditional economic data with textual information from a daily business newspaper, aiming to offer a more nuanced understanding of why the index changes over time. The author uses a daily business newspaper as a primary data source. The newspaper's content is processed and decomposed into time series, representing various news topics using a Latent Dirichlet Allocation (LDA) model. This allows the author to capture the themes and topics prevalent in the newspaper's coverage over time. The daily business cycle index is estimated using a Bayesian time-varying Dynamic Factor Model (DFM) that combines observed daily and quarterly economic data. The model enforces dynamic sparsity on the factor loadings, allowing for a more flexible and robust representation of the business cycle. The newly constructed index, known as the Newsy Coincident Index (NCI), outperforms traditional business cycle indicators in terms of timeliness and accuracy. It also provides valuable insights into the specific news topics that influence changes in the business cycle index.

(van Binsbergen et al., 2022a) construct a 170-year-long time series measure of economic sentiment at both the national and state levels using text data from 13,000 US local newspapers comprising about 1 billion articles. They employ state-of-the-art machine learning methods to create this extensive dataset, providing an expanded perspective on economic sentiment in both the time series (covering over a century)

and the cross-section. To measure economic sentiment, they customize a state-of-the-art machine learning technique. This approach generates a continuous measure of sentiment for each word and phrase in the dictionary, allowing for nuanced sentiment analysis. The economic sentiment measure exhibits strong predictive power for various economic fundamentals, including GDP growth, consumption, and employment growth. It even leads consensus GDP forecasts, indicating its relevance for forecasting. The predictive power of their measure primarily stems from forward-looking sentiment related to future economic events. The economic sentiment measure mainly operates through the labor channel, affecting employment, consumption, and services rather than the capital channel. It does not predict inflation.

Traditional news is now complemented by social media news, with more people consuming news on social media platforms such as Facebook, Instagram, Twitter (now X), or TikTok. (Azar and Lo, 2016) discuss the use of social media data, particularly tweets related to the Federal Reserve (FOMC) meetings, in predicting asset prices and constructing trading strategies. Traditional methods for measuring investor sentiment have limitations, and social media provides a new data source. The study uses natural language processing to assign polarity scores to tweets and analyzes their impact on asset returns. They find that tweets related to the Federal Reserve can predict returns of the CRSP Value-Weighted Index, even when considering tweets published at least 24 hours before FOMC meetings. Trading strategies based on Twitter sentiment can outperform passive buy-and-hold strategies as well as a comparable dynamic asset allocation strategy that does not use Twitter information. Data was collected from 2007 to 2014, and the analysis suggests that tweets contain valuable information for asset pricing during that period. The study explores various regression models to predict asset returns, showing that Twitter sentiment is particularly influential on days when the FOMC meets. The study discusses investment strategies based on the Kelly Criterion, demonstrating that portfolios using Twitter data can outperform benchmark portfolios, especially with higher leverage. The results suggest that tweets can be a valuable source of information for investors, even when considering the limitations of social media data.

10.5.3 Firm-Level Sentiment Analysis

10-K annual reports are critical to disclose financial and corporate information about public companies to the investment community. How are they evolving over time? (Dyer et al., 2017) investigate trends in the textual content of these reports filed by companies from 1996 to 2013. They analyze various aspects of 10-K disclosures, such as length, readability, boilerplate language, redundancy, specificity, and the prevalence

of hard information. They use a natural language processing technique called Latent Dirichlet Allocation (LDA) to understand the changing content of these reports and identify underlying topics. The study documents several trends in the textual characteristics of 10-K disclosures over the examined period. These trends include an increase in the length of 10-Ks, greater use of boilerplate language, increased redundancy, reduced readability, lower specificity, and a decreased proportion of hard information in the reports.

How does context influence the informativeness of earnings figures, and do financial analysts incorporate contextual information in their forecasts? (Kim and Nikolaev, 2023) aim to answer these questions. They utilize deep learning techniques, including BERT, to quantify the textual context related to disclosed earnings in a comprehensive dataset spanning from 1995 to 2021. They create different models to assess the impact of narrative context on interpreting financial data, including models that use only numeric inputs, only contextual inputs, and models that allow context to influence numeric interpretation. They find accounting information and contextual information is similarly informative when used independently. Combining both numeric and contextual inputs significantly improves the accuracy of predictions about future earnings changes, cash flows, and stock returns.

There is a noticeable increasing trend in the importance of context for interpreting accounting numbers over time. Analysts' forecast revisions are influenced by the earnings context disclosed in MD&A statements. Earnings context helps identify heterogeneity in earnings persistence, which is an indicator of earnings quality.

The SEC Form 8-K filings contain textual information about significant corporate events and activities, making them valuable for predicting stock returns and other financial insights. (Han et al., 2022b) use them to forecast stock excess returns through machine learning techniques. They model the stock excess return forecasting as an imbalanced multiclass classification problem. They use daily excess returns as the target variable and discretize them into three classes. This approach addresses the imbalanced nature of the problem, where one class has significantly more observations than others. A multiclass Support Vector Machines (SVM) model is used for stock excess return forecasting. This model outperforms other machine learning and deep learning methods, demonstrating its efficiency and reliability. The study compares different text vectorization models, including Bidirectional Encoder Representations from Transformers (BERT), term frequency-inverse document frequency ($tf - idf$), and Bag-of-words (BoW). $tf - idf$ vectorization is found to outperform BERT in stock excess return forecasting.

Are pre-existing sentiment dictionaries important to analyze financial documents? (Ke et al., 2020) propose a novel model-based approach to sentiment analysis without relying on them. The proposed approach, called SESTM (Sentiment Extraction via Screening and Topic Modeling), consists of three main steps:

- Screening for Sentiment-Charged Words: The method identifies words that carry sentiment information by calculating how frequently they co-occur with positive returns in financial news articles.
- Learning Sentiment Topics: After identifying sentiment-charged words, the model estimates the positive and negative sentiment topics using a supervised learning approach based on the relationship between word frequencies and stock returns.
- Scoring New Articles: Finally, the method allows for the estimation of sentiment scores for new articles by maximizing the likelihood of the observed word frequencies and returns, with a penalty term to account for the limited data and low signal-to-noise ratio.

The advantages of SESTM include its simplicity, minimal computing power requirements, and the ability to create a sentiment scoring model customized for specific data contexts. The study presents empirical analyses demonstrating the model's predictive capacity in the finance domain and its superiority over existing sentiment analysis methods such as the LM model (Loughran McDonald model), the commercial solution called Ravenpack, and power words (Jegadeesh and Wu, 2013). In particular, it documents trading strategies that go long a company when the sentiment is positive and short if it is negative that generate good returns and Sharpe ratios.

Sentiment analysis can be applied to social media. (Bartov et al., 2018) explore the potential of Twitter (now called X) in predicting firm-level earnings and stock returns. They investigate whether individual opinions expressed on Twitter just before a company's earnings announcement can forecast its earnings and stock price reactions. The study covers the period from 2009 to 2012 and examines a vast sample of tweets related to publicly traded companies. The study finds that the aggregate opinion derived from individual tweets about a company can successfully predict the firm's upcoming quarterly earnings. This predictive ability holds true for tweets that convey both original and existing information. The research also reveals a positive association between the immediate stock price reaction to quarterly earnings announcements and the aggregate Twitter opinion. This relationship remains consistent when tweets are categorized into those conveying original information and those disseminating existing information. Twitter serves a dual role in the capital market by providing both original information

about a company's prospects and disseminating existing information. Tweets related to earnings, firm fundamentals, and stock trading play a more crucial role in predicting earnings, but both types of tweets have a similar association with announcement returns. The predictive power of Twitter opinions is stronger for companies operating in weaker information environments, emphasizing its significance as a source of information in such settings.

10.5.4 Risk Management and Sentiment-driven Indicators

The 10-K Management Discussion & Analysis (MD&A) section is where management discusses and analyzes the performance of the company in greater detail. (Hoberg and Lewis, 2017) explore the use of text-based analysis (Latent Dirichlet Allocation (LDA)) of the 10-K MD&A disclosures to identify fraudulent firms. They find that fraudulent firms tend to produce abnormal verbal disclosures compared to non-fraudulent counterparts: Fraudulent firms disclose fewer details explaining their performance, firms engaged in revenue fraud tend to emphasize revenue growth in their disclosures, and firms involved in expense fraud tend to tell more about research and development, potentially to inflate perceived growth options, fraudulent firms also tend to under-disclose information related to liquidity challenges, fraudulent firms also reveal more information related to acquisitions, indicating potential motives linked to acquisition pricing. (Brown et al., 2020) also find that topics discussed in annual report filings and the attention devoted to each topic are useful signals in detecting financial misreporting.

In credit markets, (Wang, 2020) investigates whether the language tone used in corporate risk-related textual disclosures provides valuable information to the credit default swap (CDS) market. Using a substantial dataset of textual risk disclosures found in the Management's Discussion and Analysis (MD&A) sections of 10-K and 10-Q filings, the author finds that changes in CDS spreads around the 10-K/Q filing date are positively associated with the pessimism of the language used in risk disclosures. Cross-sectional analyses reveal that the CDS market's response to the tone of MD&A risk disclosures is more pronounced for reference entities that are closer to default, suggesting that creditors are particularly concerned about downside risk. Additionally, the CDS market reacts more significantly to the tone of MD&A risk disclosures for reference entities operating in weaker information environments.

In addition to financial disclosures, earnings conference calls could also be informative about risk. (Larcker and Zakolyukina, 2012) look at the detection of deceptive discussions in quarterly earnings conference calls, with a particular emphasis on the narratives of CEOs and CFOs. They want to predict deceptive financial reporting by

analyzing the language used in these executive statements. The research uses transcripts of conference calls collected by FactSet Research Systems Inc. and identifies financial restatements. Four different criteria are used to label a conference call as "deceptive," including material weaknesses, auditor changes, late filings, irregularities, and SEC investigations. The classification models based on CFO and CEO narratives outperform random guessing by 6% - 16% in out-of-sample tests. These linguistic models perform statistically as well as or better than accounting models that rely on discretionary accruals and other accounting variables. Deceptive CEOs and CFOs exhibit linguistic patterns distinct from truthful executives, including more references to general knowledge, fewer non-extreme positive emotion words, and fewer references to shareholder value. Deceptive CEOs use more extreme positive emotion words and fewer anxiety words and use more negation and extremely negative emotion words.

Given that companies' disclosures are receiving more scrutiny from algorithms, we expect management to react and create disclosures that are more machine-friendly. In this context, (Cao et al., 2023) explore how companies adapt their corporate disclosures in response to the growing presence of machine readers and artificial intelligence (AI) equipped investors. They find that post-2011 (the publication year of a specialized finance dictionary), firms expecting high machine downloads tend to avoid using negative words from the finance dictionary, suggesting a shift towards more positive sentiment in their disclosures.

Not only firms' financial disclosures but also analysts' reports can have an impact on stock returns. (van Binsbergen et al., 2022b) focus on the textual analysis of short-seller research reports and their impact on targeted firms. They analyze the content of these reports, with a specific emphasis on allegations of accounting fraud and earnings mismanagement. The study reveals that a significant portion of short-seller research reports deals with allegations of accounting fraud and earnings mismanagement. Approximately 95% of these reports mention fraud-related words, and around 80% of the report text is dedicated to these allegations. The study suggests that investors tend to underreact to the information contained in short-seller reports, particularly regarding long-term cash flows and real investment implications. On average, targeted firms experience abnormal returns of -4.9% on the publication day, and prices continue to decline over 12 months, resulting in an average cumulative negative effect of -15%. Such reports are found to have significant real economic consequences, including reduced corporate investment and stock issuances for targeted companies.

10.6 Case Studies

10.6.1 Disclosure Sentiment: Machine Learning vs Dictionary Methods

(Frankel et al., 2022) revisit the measure of disclosure sentiment (positive, negative, neutral, difference between positive and negative or net tone) in 10-K filing and conference call updates. The study is based on previous research by (Loughran and McDonald, 2011), who developed a disclosure sentiment measure using a dictionary (LM dictionary) and found it more associated with returns at 10-K filing dates than a Harvard Psychosociological dictionary-based measure. The contemporaneous returns are used to validate the sentiment measures (they need to be of the same sign).

In a dictionary method, each word is assigned a sentiment tone (positive, negative, neutral), and a document receives a positive tone, negative tone, and net tone variables based on the average of the word's tones. Words like "abandon," "abdicate," and "flaw" are negative, while "able," "abundance," and "leadership" are positive, according to the Loughran and McDonald dictionary.

However, there are limitations to dictionary methods, such as not accounting for language changes over time, industry-specific variations, inability to weigh word importance, updating costs, and potential researcher subjectivity leading to bias, while machine learning models can use rolling training data to adapt to change.

The authors use three machine learning methods: random forest regression trees (RF), support vector regression (SVR), and supervised Latent Dirichlet Allocation (sLDA) for sentiment analysis. These methods associate the words in the documents (one and two-word phrases) with the daily stock returns (the dependent variables) around the release date of the documents.

Random forest regression trees (RF) are an ensemble learning method used for regression. They work by creating multiple decision trees during training. The output is the mean prediction of the individual trees. Random forests are effective in reducing overfitting compared to single-decision trees.

Support vector regressions are based on support vector machines. They minimize a cost function while ensuring that the predictions for each training sample are within a specified threshold range of the true values. The model is trained by adjusting parameters, such as weights and bias, to achieve this objective.

Supervised Latent Dirichlet Allocation (Blei and McAuliffe, 2007) groups words in documents into topics that can help predict an objective. The sLDA model jointly models documents and responses, accommodating different types of response variables. It draws topic proportions, topic assignments, words, and response variables. The model parameters use variational expectation-maximization (EM) procedures.

They collected 74,363 firm-year observations of 10-K documents from 1996 to 2019. Every year, they train each model on one year of data and predict the document sentiment for the following year. With the sentiment score, they run the regression below to see how it correlates with the stock return surrounding the 10-K filing date:

$$\begin{aligned} &10K\ CAR[0,1]_{i,t} \\ &= \alpha_0 + \alpha_1 SENTIMENT_{i,t} + \alpha_2 ln(MVE_{i,t}) + \alpha_3 BTM_{i,t} \\ &+ \alpha_4 TURNOVER_{i,t} + \alpha_5 PREFFALPHA_{i,t} + \alpha_6 INSTOWN_{i,t} \\ &+ \alpha_7 NASDAQ_{i,t} + \epsilon_{i,t} \end{aligned}$$

- $10K\ CAR[0,1]_{i,t}$ is the cumulative market-adjusted return for the firm in the [0,1] trading window surrounding the current-year 10-K filing date.
- $BTM_{i,t}$ is the book-to-market ratio calculated as the firm's book value of common equity at the end of the quarter or year divided by MVE.
- $TURNOVER_{i,t}$ is the number of shares traded for the firm in the trading days [-252,-6] relative to the conference call or 10-K filing date divided by the firm's shares outstanding at the conference call or 10-K filing date.
- $PREFFALPHA_{i,t}$ is the Fama–French alpha based on a regression of their three-factor model using trading days [-252, -6] relative to the conference call or 10-K filing date. At least 60 observations of daily returns must be available to be included in the sample.
- $INSTOWN_{i,t}$ is the percentage of shares of the firm held by institutional investors.
- $NASDAQ_{i,t}$ is an indicator variable equal to the 1 if the firm is traded on the NASDAQ, and equal to 0 otherwise.

They find that the RF model using its sentiment indicator gives the regression with the highest R2 compared to the dictionary approach and the other two machine learning models.

They perform a similar analysis with conference call texts. Their data on conference calls contains 106,151 firm-quarter conference call transcripts between 2004 and 2016. They run the following regression:

$$\begin{aligned} & CC\ CAR[0,1]_{i,q} \\ &= \alpha_0 + \alpha_1 SENTIMENT_{i,q} + \alpha_2 EARNSURP_{i,q} + \alpha_3 ln(MVE_{i,q}) \\ &+ \alpha_4 BTM_{i,q} + \alpha_5 TURNOVER_{i,q} + \alpha_6 PREFFALPHA_{i,q} \\ &+ \alpha_7 INSTOWN_{i,q} + \alpha_8 NASDAQ_{i,q} + \epsilon_{i,qt} \end{aligned}$$

- $EARNSURP_{i,q}$ is the earnings per share for the firm in the current quarter less the mean earnings per share forecast for the firm made prior to the current-quarter earnings announcement date scaled by the firm's stock price at the end of the quarter. Each analyst uses the latest forecast before the current-quarter earnings announcement date, removing any forecasts made more than 90 days before the earnings announcement date.
- $MVE_{i,q}$ is the market value of equity for the firm in the current quarter or year calculated as the firm's stock price multiplied by the number of shares outstanding at the end of the quarter or year.

The results indicate that the machine learning models, and in particular the RF model, consistently capture sentiment in both 10-K filings and conference calls.

They also show that their sentiment measures seem to predict earnings surprises and, therefore, that they contain information overlooked by investors. The study concludes that machine-learning methods offer a more robust and reliable approach to measuring disclosure sentiment compared to dictionary-based methods.

10.6.2 FinBERT: Financial Sentiment Analysis with Pre-trained Language Models

(Araci, 2019) proposes FinBERT, a language model based on BERT applied to financial sentiment analysis. The model is tested on two benchmark datasets, FiQA

("FiQA - 2018," n.d.) and Financial PhraseBank (Malo et al., 2013), and is compared to three baseline models: an LSTM model with Glove embeddings, an LSTM model with ELMo embeddings and the ULMFit model (Universal Language Model Fine-tuning) from (Howard and Ruder, 2018).

The FiQA benchmark was introduced for an open challenge that took place at WWW 2018 in Lyon, France, focusing on Natural Language Processing (NLP) techniques applied to the financial domain. The challenge aimed to advance aspect-based sentiment analysis and opinion-based Question Answering for financial data. Two tasks were presented: aspect-based financial sentiment analysis and opinion-based Question Answering over financial data. There are 1,174 news headlines and tweets with their sentiment scores.

Financial PhraseBank is a phrase bank for sentiment analysis of financial news related to companies listed on OMX Helsinki. Ten thousand articles covering various types of companies, industries, and news sources were collected using automated web scraping. Sentences containing no units with a perceptible semantic orientation were excluded, leaving 53,400 sentences. A random subset of approximately 5,000 sentences was selected for annotation. The annotation task involved classifying each sentence as positive, negative, or neutral from an investor's perspective, considering the potential impact on stock prices. 28.1% of the sentences are positive, 12.4% are negative, and 59.4% are neutral. They are divided into 60% for training, 20% for testing, and 20% for validation.

FinBERT is developed in three steps:

- BERT is further pre-trained to predict the following sentences on a financial corpus TRC2-financial, a subset of the Thomson Reuters Text Research Collection (TRC2) ("Reuters Corpora @ NIST," n.d.)
- For classification, a dense layer replaces the next sentence prediction layer from the pre-trained model. The weights from the pretraining of the layers preceding the dense layer are preserved.
- It is fine-tuned to avoid catastrophic forgetting by following (Howard and Ruder, 2018) and using slanted triangular learning rates (increasing then decreasing learning rate), discriminative fine-tuning (lower learning rates for lower layers), and gradual unfreezing (layers are unfrozen from the top layer in training).

FinBERT is compared to three baseline methods: an LSTM classifier with GLoVe embeddings, an LSTM classifier with ELMo embeddings, and a ULMFit classifier.

ULMFit (Howard and Ruder, 2018) is a Universal Language Model Fine-tuning that allows for sample-efficient transfer learning and uses techniques like discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing.

FinBERT is evaluated on the Financial PhraseBank dataset according to loss, accuracy, and F1 score and is shown to outperform the three baseline methods with a lower loss, a higher accuracy (15% higher than the state-of-the-art), and a higher F1 score.

Further pre-training BERT on the specialized domain effectively lowers the loss and improves the accuracy in a 10-fold cross-validation.

Using the three methods suggested by (Howard and Ruder, 2018) is effective at preventing catastrophic forgetting prevention. Furthermore, it is not necessary to retrain the whole BERT model; instead, the last layer for classification just needs to be retrained. There is no gain in performance in retraining beyond the ninth layer.

10.6.3 Can ChatGPT Decipher Fedspeak?

(Hansen and Kazinnik, 2023) discuss the empirical evaluation of Generative Pre-trained Transformer (GPT) models, especially ChatGPT, in their ability to understand and classify the language used by the Federal Reserve, known as Fedspeak. Fedspeak is often convoluted and hard to decipher.

They use GPT-3 to classify FOMC statements into five categories: dovish, mostly dovish, neutral, mostly hawkish, and hawkish, depending on the policy stance of the FOMC towards simulating or slowing the economy. They describe the categories as follows:

- *Dovish (-1): Strongly expresses a belief that the economy may be growing too slowly and may need stimulus through monetary policy.*
- *Mostly dovish (-0.5): Overall message expresses a belief that the economy may be growing too slowly and may need stimulus through monetary policy.*
- *Neutral (0): Expresses neither a hawkish nor dovish view and is mostly objective.*
- *Mostly hawkish (0.5): Overall message expresses a belief that the economy is growing too quickly and may need to be slowed down through monetary policy.*
- *Hawkish (1): Strongly expresses a belief that the economy is growing too quickly and may need to be slowed down through monetary policy.*

In Figure 10.13, we feed a recent FOMC (Figure 10.12) statement to ChatGPT and ask it to classify the sentences. The results are in Figure 10.14.

February 01, 2023

Federal Reserve issues FOMC statement

For release at 2:00 p.m. EST

Recent indicators point to modest growth in spending and production. Job gains have been robust in recent months, and the unemployment rate has remained low. Inflation has eased somewhat but remains elevated.

Russia's war against Ukraine is causing tremendous human and economic hardship, contributing to elevated global uncertainty. The Committee is highly attentive to inflation risks.

The Committee seeks to achieve maximum employment and inflation at the rate of 2 percent over the longer run. In support of these goals, the Committee decided to raise the target range for the federal funds rate to 4-1/2 to 4-3/4 percent. The Committee anticipates that ongoing increases in the target range will be appropriate in order to attain a stance of monetary policy that is sufficiently restrictive to return inflation to 2 percent over time. In determining the extent of future increases in the target range, the Committee will take into account the cumulative tightening of monetary policy, the lags with which monetary policy affects economic activity and inflation, and economic and financial developments. In addition, the Committee will continue reducing its holdings of Treasury securities and agency debt and agency mortgage-backed securities, as described in its previously announced plans. The Committee is strongly committed to returning inflation to its 2 percent objective.

In assessing the appropriate stance of monetary policy, the Committee will continue to monitor the implications of incoming information for the economic outlook. The Committee would be prepared to adjust the stance of monetary policy as appropriate if risks emerge that could impede the attainment of the Committee's goals. The Committee's assessments will take into account a wide range of information, including readings on labor market conditions, inflation pressures and inflation expectations, and financial and international developments.

Voting for the monetary policy action were Jerome H. Powell, Chair; John C. Williams, Vice Chair; Michael S. Barr; Michelle W. Bowman; Lael Brainard; Lisa D. Cook; Austan D. Goolsbee; Patrick Harker; Philip N. Jefferson; Neel Kashkari; Lorie K. Logan; and Christopher J. Waller.

Figure 10.12. FOMC Statement of February 1, 2023. Source: https://www.federalreserve.gov/newsevents/pressreleases/monetary20230201a.htm

ChatGPT's answers are reasonable except for sentence 9, which states "The Committee is strongly committed to returning inflation to its 2 percent objective", which is more of a hawkish statement than a dovish one. ChatGPT might have confused the interest rate (where low is dovish) with the inflation rate (interest rates might need to be high for inflation to come down to 2%).

Nonetheless, GPT-3 classification (in a zero-shot version and fine-tuned version) has been compared to three vocabulary-based models, including the Loughran-McDonald model, and was shown to outperform along these different metrics mostly:

- Mean absolute error (MAE)
- Root mean squared errors (RMSE)
- Accuracy, the proportion of correct predictions
- Kappa, the agreement between predictions and actual values accounting for the chance of agreement
- F1 score, the harmonic mean of precision and recall
- Balanced accuracy, the average of true positive predictions among all positive cases and true negative predictions among all negative cases

Imagine you are a research assistant working for the Fed. You have a degree in Economics. Your task is to classify the following five sentences into one of the five categories (dovish, mostly dovish, neutral, mostly hawkish, hawkish). These sentences are taken at random from the texts of FOMC announcements:

1. Recent indicators point to modest growth in spending and production. Job gains have been robust in recent months, and the unemployment rate has remained low.
2. Inflation has eased somewhat but remains elevated.
3. Russia's war against Ukraine is causing tremendous human and economic hardship and is contributing to elevated global uncertainty.
4. The Committee is highly attentive to inflation risks.

5. The Committee seeks to achieve maximum employment and inflation at the rate of 2 percent over the longer run. In support of these goals, the Committee decided to raise the target range for the federal funds rate to 4-1/2 to 4-3/4 percent.
6. The Committee anticipates that ongoing increases in the target range will be appropriate in order to attain a stance of monetary policy that is sufficiently restrictive to return inflation to 2 percent over time.
7. In determining the extent of future increases in the target range, the Committee will take into account the cumulative tightening of monetary policy, the lags with which monetary policy affects economic activity and inflation, and economic and financial developments.
8. In addition, the Committee will continue reducing its holdings of Treasury securities and agency debt and agency mortgage-backed securities, as described in its previously announced plans.
9. The Committee is strongly committed to returning inflation to its 2 percent objective.
10. In assessing the appropriate stance of monetary policy, the Committee will continue to monitor the implications of incoming information for the economic outlook.
11. The Committee would be prepared to adjust the stance of monetary policy as appropriate if risks emerge that could impede the attainment of the Committee's goals.
12. The Committee's assessments will take into account a wide range of information, including readings on labor market conditions, inflation pressures and inflation expectations, and financial and international developments.

Figure 10.13. Prompt to ChatGPT

Furthermore, GPT models can provide explanations and reasoning for their classifications, aligning with human reasoning in many cases. This capability has implications for automating the analysis of Fed communications, potentially improving transparency and effectiveness in policy communication.

Here is the classification of the sentences into the five categories (dovish, mostly dovish, neutral, mostly hawkish, hawkish) based on the tone and content of the sentences:

1. Mostly Dovish: This sentence mentions modest growth and robust job gains, which are positive economic indicators.
2. Neutral: This sentence mentions elevated inflation but does not provide a clear indication of a policy stance.
3. Neutral: This sentence discusses global uncertainty due to Russia's war against Ukraine but does not directly address monetary policy.
4. Neutral: This sentence indicates that the Committee is attentive to inflation risks but does not explicitly state a policy stance.
5. Hawkish: This sentence announces an increase in the federal funds rate, which is a hawkish move.
6. Hawkish: This sentence suggests that ongoing rate increases are expected to achieve a restrictive monetary policy stance to combat inflation.
7. Mostly Hawkish: This sentence implies a willingness to tighten monetary policy further if necessary, indicating a hawkish stance.
8. Mostly Hawkish: This sentence mentions the reduction of holdings of securities, which is generally a hawkish move.
9. **Mostly Dovish: This sentence emphasizes the Committee's commitment to returning inflation to its 2 percent objective, indicating a dovish stance.**
10. Neutral: This sentence mentions monitoring incoming information for economic outlook but does not express a clear policy stance.
11. Neutral: This sentence indicates a willingness to adjust policy if risks emerge but does not specify a particular direction.
12. Neutral: This sentence mentions factors the Committee considers but does not express a clear policy stance.

Figure 10.14. Classification and Explanations from ChatGPT

The research also explores how GPT models can identify macroeconomic shocks using the narrative approach (Romer and Romer, 1989; Romer and Romer, 2022). This approach involves reading and inferring causal relationships from relevant texts, and GPT models show promise in this area as well after being given the definition of a policy shock (Figure 10.15). This is a significant benefit as this work was done manually by researchers and, therefore, was time-consuming and not scalable.

As a monetary policy expert, your task is to determine whether a given text contains a monetary policy shock. A monetary policy shock refers to movements in monetary policy that are unrelated to current or prospective real economic activity. These shocks occur when policy makers change money growth and interest rates due to concerns about prevailing inflation levels, even when the economy is stable. Policy makers, in these instances, are willing to accept potential negative consequences for aggregate output and unemployment.

Analyzing the provided text, determine whether it meets the criteria for a monetary policy shock based on the following factors:

- The policy makers believed the economy was at potential output.
- Policy makers changed money growth and interest rates due to high inflation.
- Policy makers understood and accepted the potential adverse consequences for output and unemployment.

provided text:

Participants' Views on Current Conditions and the Economic Outlook

In their discussion of current economic conditions, participants noted that economic activity had been expanding at a moderate pace. Job gains had been robust in recent months, and the unemployment rate remained low. Inflation remained elevated. Participants agreed that the U.S. banking system was sound and resilient. They commented that tighter credit conditions for households and businesses were likely to weigh on economic activity, hiring, and inflation. However, participants agreed that the extent of these effects remained uncertain. Against this background, the Committee remained highly attentive to inflation risks.

In assessing the economic outlook, participants noted that real GDP growth had continued to exhibit resilience in the first half of the year and that the economy had been showing considerable momentum. A gradual slowdown in economic activity nevertheless appeared to be in progress, consistent with the restraint placed on

demand by the cumulative tightening of monetary policy since early last year and the associated effects on financial conditions. Participants remarked on the uncertainty about the lags in the effects of monetary policy on the economy and discussed the extent to which the effects on the economy stemming from the tightening that the Committee had undertaken had already materialized. Participants commented that monetary policy tightening appeared to be working broadly as intended and that a continued gradual slowing in real GDP growth would help reduce demand–supply imbalances in the economy. Participants assessed that the ongoing tightening of credit conditions in the banking sector, as evidenced in the most recent surveys of banks, also would likely weigh on economic activity in coming quarters. Participants noted the recent reduction in total and core inflation rates. However, they stressed that inflation remained unacceptably high and that further evidence would be required for them to be confident that inflation was clearly on a path toward the Committee's 2 percent objective. Participants continued to view a period of below-trend growth in real GDP and some softening in labor market conditions as needed to bring aggregate supply and aggregate demand into better balance and reduce inflation pressures sufficiently to return inflation to 2 percent over time.

Participants noted that consumer spending had recently exhibited considerable resilience, underpinned by, in aggregate, strong household balance sheets, robust job and income gains, a low unemployment rate, and rising consumer confidence. Nevertheless, tight financial conditions, primarily reflecting the cumulative effect of the Committee's shift to a restrictive policy stance, were expected to contribute to slower growth in consumption in the period ahead. Participants cited other factors that were likely leading to, or appeared consistent with, a slowdown in consumption, including the declining stock of excess savings, softening labor market conditions, and increased price sensitivity on the part of customers. Some participants observed that recent increases in home prices suggested that the housing sector's response to monetary policy restraint may have peaked.

In their discussion of the business sector, participants cited various improvements in firms' cost structures. These included better-functioning supply chains, lower input costs, and an increased ability to hire and retain workers. Participants also discussed conditions that could lead to higher economic activity—such as leaner inventories and reduced expectations of a sharp economic slowdown—and factors that could lead to lower economic activity—such as continuing economic uncertainty, the vulnerabilities of the CRE market, and the ongoing weakness of manufacturing output. Participants judged that, over coming quarters, firms would

reduce the pace of their investment spending and hiring in response to tight financial conditions and the slowing of economic activity.

Participants remarked that the labor market continued to be very tight but pointed to signs that demand and supply were coming into better balance. They noted evidence that labor demand was easing—including declines in job openings, lower quits rates, more part-time work, slower growth in hours worked, higher unemployment insurance claims, and more moderate rates of nominal wage growth. In addition, they remarked on indications of increasing labor supply, including a further rise in the prime-age participation rate to a post-pandemic high. Participants also observed, however, that although growth in payrolls had slowed recently, it continued to exceed values consistent over time with an unchanged unemployment rate, and that nominal wages were still rising at rates above levels assessed to be consistent with the sustained achievement of the Committee's 2 percent inflation objective. Participants judged that further progress toward a balancing of demand and supply in the labor market was needed, and they expected that additional softening in labor market conditions would take place over time.

Participants cited a number of tentative signs that inflation pressures could be abating. These signs included some softening in core goods prices, lower online prices, evidence that firms were raising prices by smaller amounts than previously, slower increases in shelter prices, and recent declines in survey estimates of shorter-term inflation expectations and of inflation uncertainty. Various participants discussed the continued stability of longer-term inflation expectations at levels consistent with 2 percent inflation over time and the role that the Committee's policy tightening had played in delivering this outcome. Nonetheless, several participants commented that significant disinflationary pressures had yet to become apparent in the prices of core services excluding housing.

Participants observed that, notwithstanding recent favorable developments, inflation remained well above the Committee's 2 percent longer-term objective and that elevated inflation was continuing to harm businesses and households—low-income families in particular. Participants stressed that the Committee would need to see more data on inflation and further signs that aggregate demand and aggregate supply were moving into better balance to be confident that inflation pressures were abating and that inflation was on course to return to 2 percent over time.

Participants generally noted a high degree of uncertainty regarding the cumulative effects on the economy of past monetary policy tightening. Participants cited upside risks to inflation, including those associated with scenarios in which recent supply

chain improvements and favorable commodity price trends did not continue or in which aggregate demand failed to slow by an amount sufficient to restore price stability over time, possibly leading to more persistent elevated inflation or an unanchoring of inflation expectations. In discussing downside risks to economic activity and inflation, participants considered the possibility that the cumulative tightening of monetary policy could lead to a sharper slowdown in the economy than expected, as well as the possibility that the effects of the tightening of bank credit conditions could prove more substantial than anticipated.

In their discussion of financial stability, participants observed that the banking system was sound and resilient and that banking stress had calmed in recent months. Participants also noted that the most recent stress-test results indicated that large banks appeared to be well positioned to withstand a severe recession. Various participants commented on risks that could affect some banks, including unrealized losses on assets resulting from rising interest rates, significant reliance on uninsured deposits, and increased funding costs. Participants also commented on risks associated with a potential sharp decline in CRE valuations that could adversely affect some banks and other financial institutions, such as insurance companies, that are heavily exposed to CRE. Several participants noted the susceptibility of some nonbank financial institutions, such as money market funds or digital asset entities, to runs or instability. In addition, several participants emphasized the need for banks to establish readiness to use Federal Reserve liquidity facilities and for the Federal Reserve to ensure its own readiness to provide liquidity during periods of stress.

In their consideration of appropriate monetary policy actions at this meeting, participants concurred that economic activity had been expanding at a moderate pace. The labor market remained very tight, with robust job gains in recent months and the unemployment rate still low, but there were continuing signs that supply and demand in the labor market were coming into better balance. Participants also noted that tighter credit conditions facing households and businesses were a source of headwinds for the economy and would likely weigh on economic activity, hiring, and inflation. However, the extent of these effects remained uncertain. Although inflation had moderated since the middle of last year, it remained well above the Committee's longer-run goal of 2 percent, and participants remained resolute in their commitment to bring inflation down to the Committee's 2 percent objective. Amid these economic conditions, almost all participants judged it appropriate to raise the target range for the federal funds rate to 5-1/4 to 5-1/2 percent at this meeting. Participants noted that this action would put the stance of monetary policy further into restrictive territory, consistent with reducing demand–supply imbalances in the

economy and helping to restore price stability. A couple of participants indicated that they favored leaving the target range for the federal funds rate unchanged or that they could have supported such a proposal. They judged that maintaining the current degree of restrictiveness at this time would likely result in further progress toward the Committee's goals while allowing the Committee time to further evaluate this progress. All participants agreed that it was appropriate to continue the process of reducing the Federal Reserve's securities holdings, as described in its previously announced Plans for Reducing the Size of the Federal Reserve's Balance Sheet. A number of participants noted that balance sheet runoff need not end when the Committee eventually begins to reduce the target range for the federal funds rate.

In discussing the policy outlook, participants continued to judge that it was critical that the stance of monetary policy be sufficiently restrictive to return inflation to the Committee's 2 percent objective over time. They noted that uncertainty about the economic outlook remained elevated and agreed that policy decisions at future meetings should depend on the totality of the incoming information and its implications for the economic outlook and inflation as well as for the balance of risks. Participants expected that the data arriving in coming months would help clarify the extent to which the disinflation process was continuing and product and labor markets were reaching a better balance between demand and supply. This information would be valuable in determining the extent of additional policy firming that may be appropriate to return inflation to 2 percent over time. Participants also emphasized the importance of communicating as clearly as possible about the Committee's data-dependent approach to policy and its firm commitment to bring inflation down to its 2 percent objective.

Participants discussed several risk-management considerations that could bear on future policy decisions. With inflation still well above the Committee's longer-run goal and the labor market remaining tight, most participants continued to see significant upside risks to inflation, which could require further tightening of monetary policy. Some participants commented that even though economic activity had been resilient and the labor market had remained strong, there continued to be downside risks to economic activity and upside risks to the unemployment rate; these included the possibility that the macroeconomic effects of the tightening in financial conditions since the beginning of last year could prove more substantial than anticipated. A number of participants judged that, with the stance of monetary policy in restrictive territory, risks to the achievement of the Committee's goals had become more two sided, and it was important that the Committee's decisions balance the risk

of an inadvertent overtightening of policy against the cost of an insufficient tightening.

Committee Policy Actions

In their discussion of monetary policy for this meeting, members agreed that economic activity had been expanding at a moderate pace. They also concurred that job gains had been robust in recent months, and the unemployment rate had remained low. Inflation had remained elevated.

Members concurred that the U.S. banking system was sound and resilient. They also agreed that tighter credit conditions for households and businesses were likely to weigh on economic activity, hiring, and inflation but that the extent of these effects was uncertain. Members also concurred that they remained highly attentive to inflation risks.

In support of the Committee's objectives to achieve maximum employment and inflation at the rate of 2 percent over the longer run, members agreed to raise the target range for the federal funds rate to 5-1/4 to 5-1/2 percent. They also agreed that they would continue to assess additional information and its implications for monetary policy. In determining the extent of additional policy firming that may be appropriate to return inflation to 2 percent over time, members concurred that they will take into account the cumulative tightening of monetary policy, the lags with which monetary policy affects economic activity and inflation, and economic and financial developments. In addition, members agreed to continue to reduce the Federal Reserve's holdings of Treasury securities and agency debt and agency mortgage-backed securities, as described in its previously announced plans. All members affirmed that they are strongly committed to returning inflation to their 2 percent objective.

Members agreed that, in assessing the appropriate stance of monetary policy, they would continue to monitor the implications of incoming information for the economic outlook. They would be prepared to adjust the stance of monetary policy as appropriate if risks emerge that could impede the attainment of the Committee's goals. Members also agreed that their assessments will take into account a wide range of information, including readings on labor market conditions, inflation pressures and inflation expectations, and financial and international developments.

At the conclusion of the discussion, the Committee voted to direct the Federal Reserve Bank of New York, until instructed otherwise, to execute transactions in the

System Open Market Account in accordance with the following domestic policy directive, for release at 2:00 p.m.:

"Effective July 27, 2023, the Federal Open Market Committee directs the Desk to:

Undertake open market operations as necessary to maintain the federal funds rate in a target range of 5-1/4 to 5-1/2 percent.

Conduct standing overnight repurchase agreement operations with a minimum bid rate of 5.5 percent and with an aggregate operation limit of $500 billion.

Conduct standing overnight reverse repurchase agreement operations at an offering rate of 5.3 percent and with a per-counterparty limit of $160 billion per day.

Roll over at auction the amount of principal payments from the Federal Reserve's holdings of Treasury securities maturing in each calendar month that exceeds a cap of $60 billion per month. Redeem Treasury coupon securities up to this monthly cap and Treasury bills to the extent that coupon principal payments are less than the monthly cap.

Reinvest into agency mortgage-backed securities (MBS) the amount of principal payments from the Federal Reserve's holdings of agency debt and agency MBS received in each calendar month that exceeds a cap of $35 billion per month.

Allow modest deviations from stated amounts for reinvestments, if needed for operational reasons.

Engage in dollar roll and coupon swap transactions as necessary to facilitate settlement of the Federal Reserve's agency MBS transactions."

The vote also encompassed approval of the statement below for release at 2:00 p.m.:

"Recent indicators suggest that economic activity has been expanding at a moderate pace. Job gains have been robust in recent months, and the unemployment rate has remained low. Inflation remains elevated.

The U.S. banking system is sound and resilient. Tighter credit conditions for households and businesses are likely to weigh on economic activity, hiring, and inflation. The extent of these effects remains uncertain. The Committee remains highly attentive to inflation risks.

The Committee seeks to achieve maximum employment and inflation at the rate of 2 percent over the longer run. In support of these goals, the Committee decided to raise the target range for the federal funds rate to 5-1/4 to 5-1/2 percent. The Committee will continue to assess additional information and its implications for

monetary policy. In determining the extent of additional policy firming that may be appropriate to return inflation to 2 percent over time, the Committee will take into account the cumulative tightening of monetary policy, the lags with which monetary policy affects economic activity and inflation, and economic and financial developments. In addition, the Committee will continue reducing its holdings of Treasury securities and agency debt and agency mortgage-backed securities, as described in its previously announced plans. The Committee is strongly committed to returning inflation to its 2 percent objective.

In assessing the appropriate stance of monetary policy, the Committee will continue to monitor the implications of incoming information for the economic outlook. The Committee would be prepared to adjust the stance of monetary policy as appropriate if risks emerge that could impede the attainment of the Committee's goals. The Committee's assessments will take into account a wide range of information, including readings on labor market conditions, inflation pressures and inflation expectations, and financial and international developments."

Figure 10.15. Prompt to ChatGPT to analyze the FOMC Minutes of 26-27 July 2023. Source: https://www.federalreserve.gov/monetarypolicy/fomcminutes20230726.htm

Based on the provided text, it does not appear to contain a monetary policy shock. Here are the reasons why it doesn't meet the criteria for a monetary policy shock:

The policy makers did not change money growth and interest rates explicitly in response to concerns about prevailing inflation levels. While they acknowledged that inflation remained elevated and was a concern, the text suggests that monetary policy decisions were made with a broader focus on economic conditions, employment, and inflation, rather than solely based on inflation concerns.

The text indicates that policy makers considered various economic factors, including economic activity, job gains, unemployment, and inflation. They also discussed the uncertainty surrounding the effects of monetary policy tightening on the economy. This suggests that policy makers were taking a balanced approach to monetary policy and not making abrupt changes unrelated to real economic activity.

The text does not mention that policy makers were willing to accept potential adverse consequences for output and unemployment in pursuit of reducing inflation.

Instead, it emphasizes the Committee's commitment to achieving its dual mandate of maximum employment and 2 percent inflation over time, without explicit acceptance of negative consequences for output and employment.

Overall, the text describes a deliberative approach to monetary policy that considers a range of economic factors and uncertainties, rather than a sudden and unrelated change in monetary policy solely to address high inflation.

Figure 10.16. Response of ChatGPT

In conclusion, GPT models, such as ChatGPT, are shown to be valuable tools for understanding and classifying complex policy language like Fedspeak. They offer the potential to enhance communication clarity, transparency, and effectiveness in central bank communication and could lead to more accessible policy messages for various audiences.