A text-based forecasting model for equity trading

Introduction

- A lot of stock price sensitive information is contained in a textual form, e.g. public announcements, press releases, news and analyst opinions

- At the moment, portfolio managers tend to restrict their analysis to the sentiment analysis task (measuring sentiment with regard to a particular stock in news and social media)

- A better approach could be trying to solve the natural language understanding task by training a model to directly predict stock returns from textual inputs

- This paper suggests a simple neural architecture, based on attention-weighted BiLSTM encoder, which allows to easily incorporate textual features into a stock return forecasting model

Constructed dataset

Example of significant development headlines.

|	Company	Headline
1	Shell	Royal Dutch Shell PLC shuts San Pablo Bay pipeline, confirms small crude spill in California - Reuters
2	Apple	S&P assigns 'AA+' rating to Apple Inc's Australian dollar-denominated senior unsecured notes
3	Volkswagen	Porsche Automobil Holding SE And Volkswagen AG Merger On Track After U.S. Suit Dropped - Reuters
4	Novartis	Novartis AG Announces Positive Results From Final Phase III Omalizumab Registration Study In Severe Form Of Chronic Skin Disease CSU

Numerical features:

- Returns of S&P500 index

- Money flow indicator (the flow of funds into and out of a security over last 14 days)

- Forward P/E ratio (stock open price divided by 12-month earnings per share guidance)

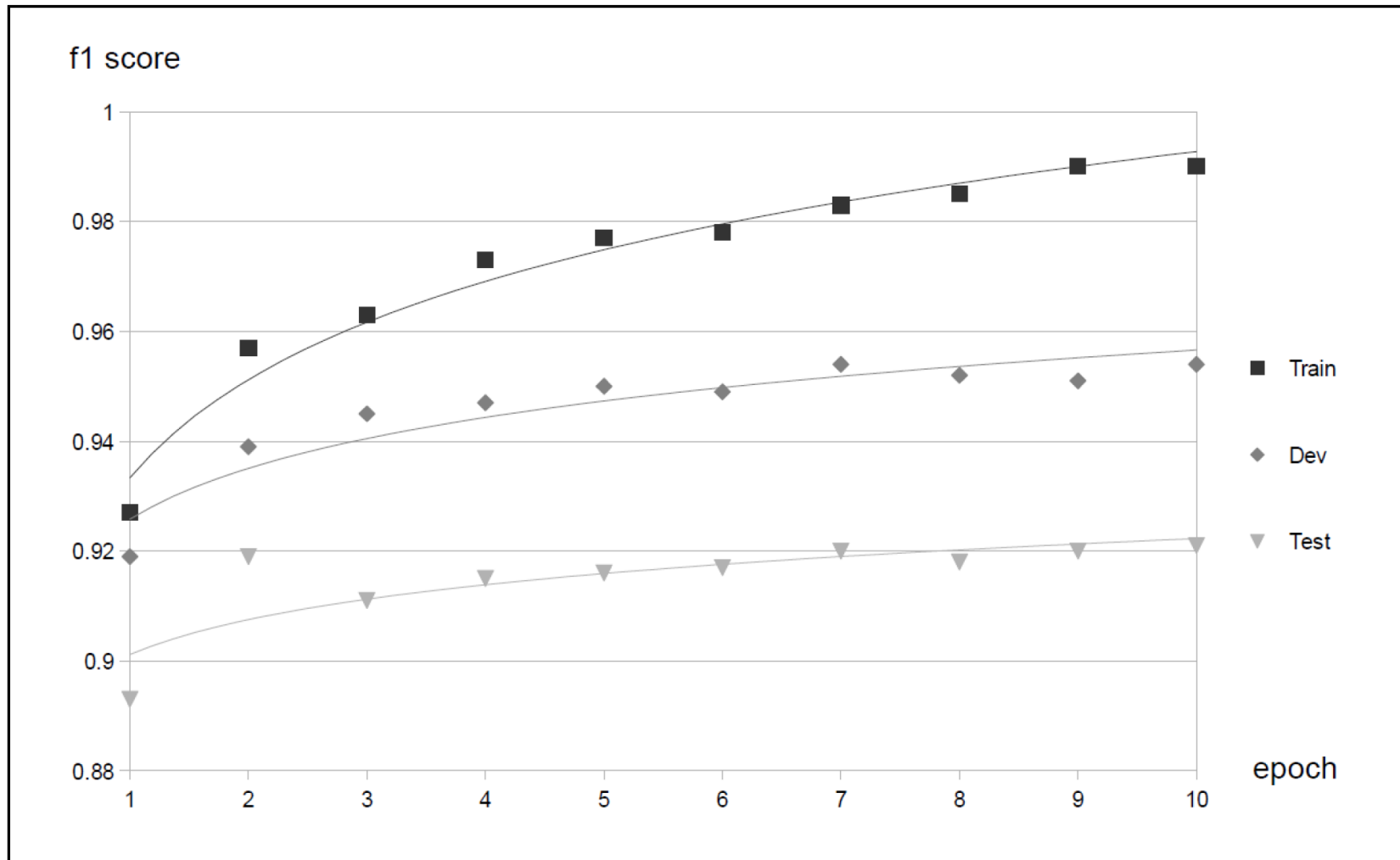- Short interest (percentage of short positions on a given stock)

Preprocessing pipeline

Filtering

Aggregation

Tokenization

PoS tagging

Lemmatization
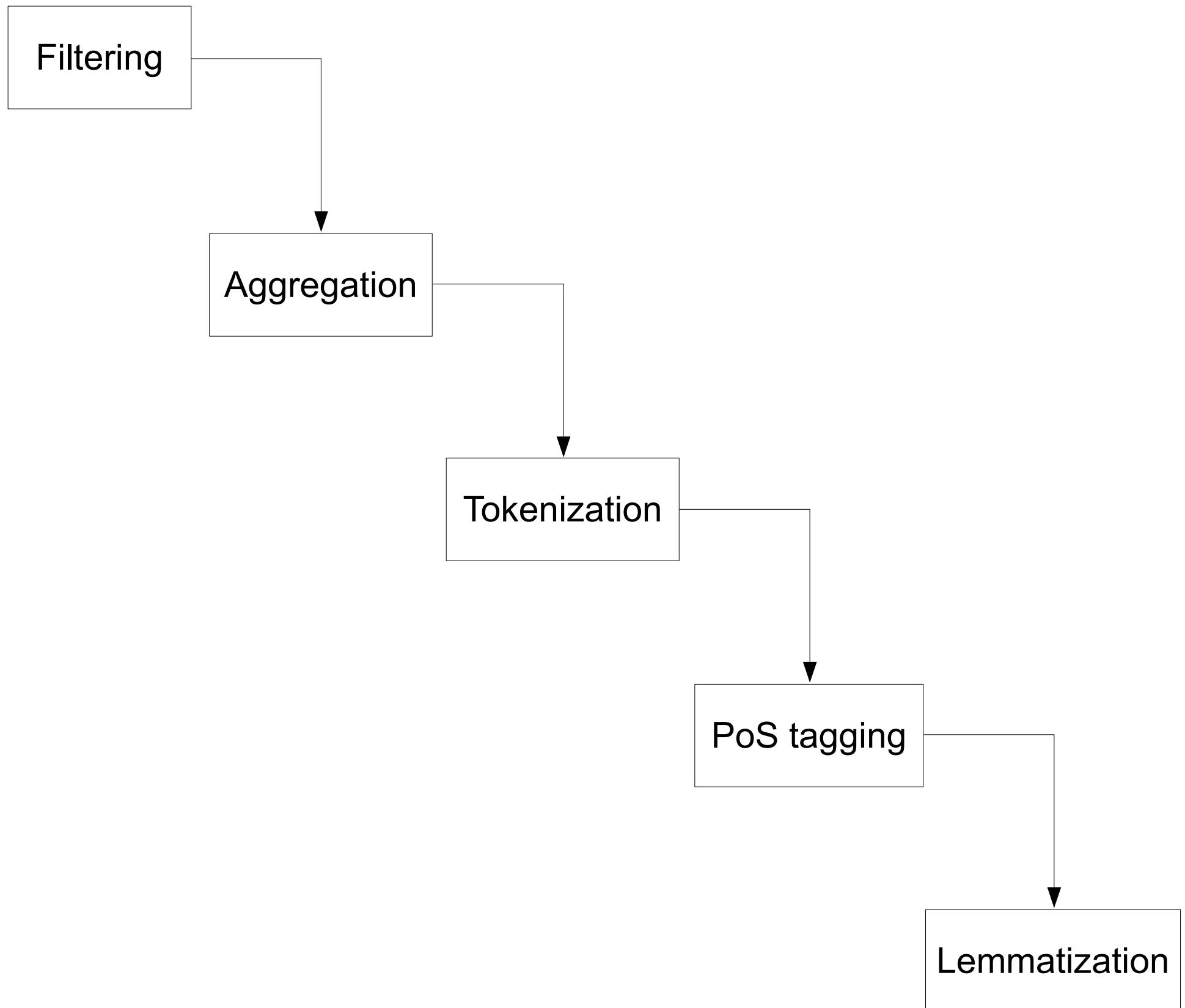
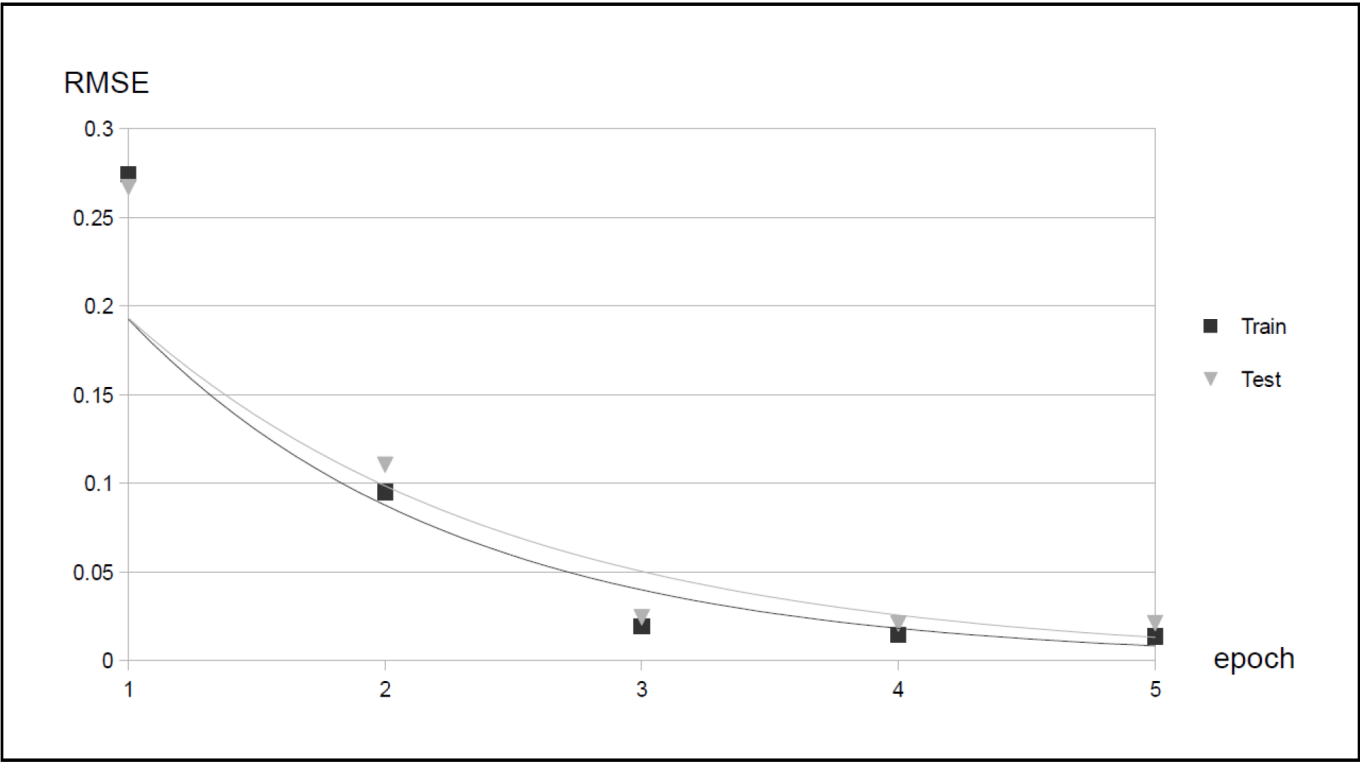# Architecture overview
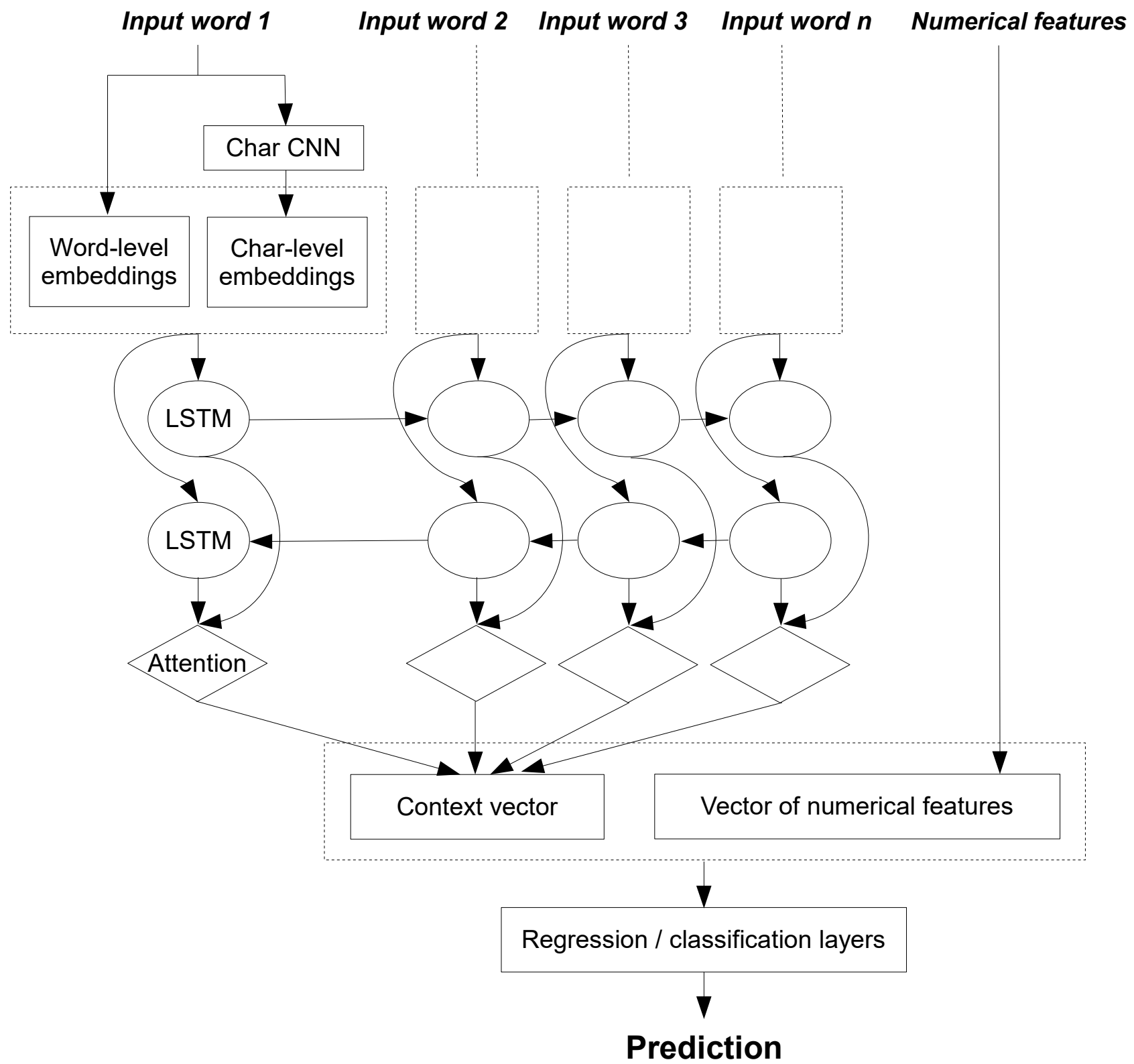
Performance on NER task (CoNLL-2003)

Learning curve on the NER task.

Performance on stock return prediction task

Learning curve on the stock return prediction task.

Evaluation of stock return forecasting models.

	Model	RMSE on train	RMSE on test
1	Neural forecasting model with textual inputs	1.37%	2.11%
2	Neural forecasting model without textual inputs	0.25%	2.36%
3	Support vector regression with tf-idf vectors	2.72%	2.98%

Future work

Possible areas of future work:

- Construction and analysis of a wider dataset of financial news and reports

- Utilization of intraday stock return data that can better capture the immediate effects of news and reports publication

- Differentiation among sources of textual data, since they may affect stock price asymmetrically

- Incorporation of deep contextualized word representations that can model polysemy, such as ELMO

- Development of a transformer forecasting architecture which can potentially improve the applied RNN-based sequence-to-sequence approach

Thank you!