

Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments

Chung-Chi Chen,¹ Hen-Hsen Huang,^{2,4} Hiroya Takamura,³ Hsin-Hsi Chen^{1,4}

¹ Department of Computer Science and Information Engineering
National Taiwan University, Taiwan

² Department of Computer Science, National Chengchi University, Taiwan

³ Artificial Intelligence Research Center,

National Institute of Advanced Industrial Science and Technology, Japan

⁴ MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
cjchen@nlg.csie.ntu.edu.tw, hhhuang@nccu.edu.tw,
takamura.hiroya@aist.go.jp, hhchen@ntu.edu.tw

Abstract

In this paper, we attempt to answer the question of whether neural network models can learn numeracy, which is the ability to predict the magnitude of a numeral at some specific position in a text description. A large benchmark dataset, called Numeracy-600K, is provided for the novel task. We explore several neural network models including CNN, GRU, BiGRU, CRNN, CNN-capsule, GRU-capsule, and BiGRU-capsule in the experiments. The results show that the BiGRU model gets the best micro-averaged F1 score of 80.16%, and the GRU-capsule model gets the best macro-averaged F1 score of 64.71%. Besides discussing the challenges through comprehensive experiments, we also present an important application scenario, i.e., detecting exaggerated information, for the task.

1 Introduction

As a prior research from a dataset obtained from Reuters, one of the largest international news agencies, over 65.66% of market comments contain numerals. Without the numerals in market comments, we will miss a lot of useful information. Table 1 lists some instances of real-time market comments. The topics include the descriptions of market data (S1), financial statements (S2), products (S3), analyst reports (S4), and events (S5). From the table, we can see that numerals provide more detailed information than words do. For example, in comment (S1) we can learn that the share price of Apple Inc. (AAPL) has fallen, but we cannot obtain the percentage change or the price quote without the numerals. Furthermore, (S3) provides crucial information such as the date (Q2) and the amount of sales with numerals (4.6

(S1) <AAPL> SHARES DOWN 4 PCT AT \$113.7 IN MORNING TRADE
(S2) <AAPL> Q1 REV VIEW \$75.08 BLN
(S3) <AAPL> - Q2 MAC SALES OF 4.6 MLN UNITS VS 4.1 MLN UNITS LAST YEAR
(S4) <AAPL>: CANACCORD GENUITY RAISES PRICE TARGET TO \$600
(S5) <AAPL> CFO SAYS REVENUE EXPECTED TO BE DOWN BETWEEN 5-10% IN CONSTANT CURRENCY FOR Q1

Table 1: Instances of market comments.

(S6) S&P 500 <.SPX> UP 1.53 POINTS AT ____ AFTER MARKET OPEN
(S7) DOW JONES <.DJI> UP 8.70 POINTS AT ____ AFTER MARKET OPEN
(S8) U.S. Q3 GDP rises ____ pct

Table 2: Instances for the proposed task.

and 4.1). These examples show the crucial roles of numerals in financial narratives.

Table 2 lists three market comments selected from our dataset as examples. Investors would know from their experiences that the blanks in (S6) and (S7) should be filled with quotes of the opening indices of the S&P 500 and Dow Jones Industrial Average (DJIA), respectively. Accordingly, they would insert a 4th-magnitude numeral, 1840, into (S6), and a 5th-magnitude numeral, 16163, into (S7). We call such an interpretation as numeracy, which is the ability to interpret simple numerical concepts at some given positions.

There are two challenging issues in (S6) and (S7): to detect the target entity, and to understand the type of information to insert into the blanks. A

more fine-grained question is shown in (S8). After getting involved in markets and reading much more news and market data, investors gain intuition about market information. For example, investors can intuitively select a 1st-magnitude numeral, 2.9, to fill in the blank in (S8). We are interested in knowing if neural network (NN) models can learn this kind of numeracy from the numerous market comments.

The contributions of this paper are four-fold: (1) providing a novel task and a benchmark dataset, called Numeracy-600K; (2) setting a strong baseline with thorough evaluation of several neural network models, including the state-of-the-art models, on the proposed task; (3) discussing the details of the challenges; and (4) indicating an important application scenario, i.e., detecting exaggerated information, for the proposed task.

The rest of this paper is organized as follows. Section 2 surveys the related work on the identification of numerals and misinformation. Section 3 defines the task and introduces the dataset used in this study. Section 4 shows and discusses the experimental results in the comprehensive experiments. Section 5 presents an application scenario of detecting exaggerated numerals in market comments. Besides, we also extend the methodology in the market comment dataset to the general article title dataset. Section 6 concludes the remarks.

2 Related Work

Murakami et al. (2017) attempted to generate market comments from stock prices. Their work used only two kinds of numerals: the latest price, and the difference of closing price between two days. As seen in Table 1, however, market comments describe various kinds of topics along with numerals. In this paper, we will provide experimental results for general market comments and show the numeracy of various NN models.

Spithourakis and Riedel (2018) used language models to predict numerals in clinical and scientific datasets. They do not touch on numeral prediction in financial market comments. In this paper, we examine whether NN models can learn numeracy to insert proper information into market comments, rather than predicting exact numerals. We will discuss the reasons in Section 4.3. Our results give a positive answer to this question.

Several different approaches have been used to detect false information and fake news. Wang

et al. (2018a) used both text information and images in tweets to detect misleading information. Tschitschek et al. (2018) identified fake news via crowd signals, namely, Facebook users flags of fake news. As mentioned in Shu et al. (2017), “the underlying characteristics of fake news have not been fully understood.” In this paper, we concentrate on market comments, and focus on exaggerated numeral identification in the comments.

3 Task Setting and Dataset

The task is defined as to test whether NN models can learn numeracy by inserting the proper magnitude of numerals into a market comment. From the human perspective, we may feel that something makes sense intuitively, but this kind of feeling is not precise. In (S9), human experience suggests that inserting 7 into the blank would be better than inserting 10. Even experienced investors may be confused, however, if the candidates are 6.9 and 7. Therefore, to test the numeracy of a model, we separate numerals into eight classes by the magnitude and ask models to predict a suitable range.

(S9) CHINA H1 GDP + _____ PCT Y/Y

For the experiments, we collected 600K market comments from Reuters. Numeracy means the approximate range of a numeral at some given position. In our task setting, we classify numerals, denoted as m , into eight classes by their magnitudes, as listed in Table 3. That is, we will examine whether NN models can insert a proper range of numerals into a market comment, rather than inserting the exact number. In addition to the eight classes, Table 3 also lists their distribution.

We predefine some extraction rules to extract the numerals in the dataset automatically. Signs (+, -, and /) were separated from numerals. Furthermore, we only considered the magnitude be-

Magnitude	Range	Ratio
Decimal	$0 \leq m < 1$	23.24
1	$1 \leq m < 10$	37.53
2	$10 \leq m < 10^2$	25.36
3	$10^2 \leq m < 10^3$	12.21
4	$10^3 \leq m < 10^4$	1.12
5	$10^4 \leq m < 10^5$	0.29
6	$10^5 \leq m < 10^6$	0.23
> 6	$10^6 \leq m$	0.01

Table 3: Distribution of numerals in the dataset.

fore the decimal point, i.e., 10.08 was classified as a 2nd magnitude. Finally, we separate the dataset into training set and test set of sizes 500k and 100k, respectively.

4 Empirical Study

4.1 Models

We adopt seven different architectures for our task, including CNN (Kim, 2014), GRU (Cho et al., 2014), BiGRU, CRNN (Choi et al., 2017), CNN-capsule (Sabour et al., 2017), GRU-capsule, and BiGRU-capsule (Wang et al., 2018b). In our models, each word in the input sentence is represented as a d -dimensional vector with word embeddings, and all the words are concatenated in as a $d \times l$ matrix, where l denotes the sentence length. Some preprocessing was performed on the data. We transformed all characters to lowercase. The sentence representation was padded to the maximum length of an instance. The target numeral to be inferred is replaced with a special token $\langle \text{TRT} \rangle$. Appendices illustrate the detailed model settings.

4.2 Experimental Results

For our task settings, each model outputs the result of the eight-way classification. We report the performance of the models in F1 scores and analyze the results by using confusion matrices.

Table 4 summarizes the experimental results. Logistic regression (LR) with bag of words, which are composed of top-1K frequent words, sets a baseline for the proposed task. The BiGRU model beats the other models with a micro-averaged F1 score of 80.16%, and the GRU-capsule model performs the best with a macro-averaged F1 score of 64.71%. The RNN-based models outperform the CNN-based models in both the general NN framework and the capsule network framework. The results account for the importance of the order of the context in market comments when inserting numerical information. Further evidence supporting this statement is that the CRNN model obtains a higher performance than the CNN model does.

Figure 1 provides the evidence for the GRU-capsule model performing the best with macro-averaged F1 score. Comparing to the other models, the GRU-capsule model correctly predicts 54% of the data in the 6th-magnitude class, which constitute 0.23% of the entire data. This result indicates that the GRU-capsule model is able to find some clues with the small size of training data.

Model	Micro-F1	Macro-F1
LR	71.25%	60.80%
CNN	77.17%	58.49%
GRU	78.25%	58.08%
BiGRU	80.16%	62.74%
CRNN	78.00%	64.62%
CNN-capsule	75.89%	59.22%
GRU-capsule	77.36%	64.71%
BiGRU-capsule	77.97%	64.34%

Table 4: Experimental results.

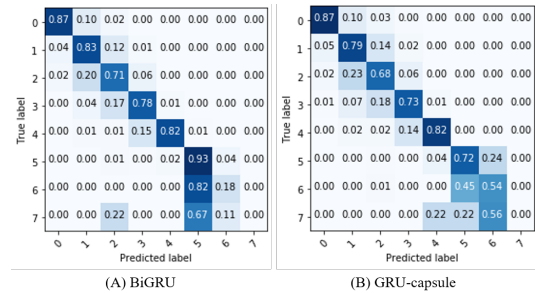


Figure 1: Confusion matrices

4.3 Error Analysis and Future Research

In this subsection, we analyze some frequent errors and point out some open issues for future research on machine learning with market comments. Table 5 lists some instances. (E1) indicates the problem of a different contract for the same financial instrument. That is, the government may publish the same bond with a different coupon rate. Whether we should replace $\langle \text{TRT} \rangle$ with 0.75 or 2.25 depends on the time of the auction described in the market comment. As another problem, the DJI in (E2) is different every day, making it hard to predict the actual amount of change. As indicated by the confusion matrix, however, the BiGRU model makes sensible predictions near the truth. (E3) shows that models should learn the past patterns (the change of the previous Disney quarterly revenue are always in 2nd magnitude) of a target company's financial statements.

(S10) *VOLVO* $\langle \text{VOLVb.ST} \rangle$: *HSBC RAISES PRICE TARGET TO SEK 105 FROM* $\langle \text{TRT} \rangle$

The numeral 10 is the ground truth for instance (E4), and 95 should be inserted into (S10), but the model predicted a 1st magnitude for (E4) and a 3rd magnitude for (S10). Both cases show that models may tend to refer to previously occurring numerals, 8 in (E4) and 105 in (S10), to decide the magnitude of the target numeral.

	T	P	Market comment	Issue
E1	0	1	CANADA <TRT> PCT 2014 BOND AUCTION YIELD LOW 1.110 PCT,HIGH 1.121 PCT	Different contract
E2	0	2	DOW JONES <.DJI> UNOFFICIALLY CLOSES UP <TRT> POINTS	Market Data
E3	1	2	Disney quarterly revenue rises <TRT> pct	Past patterns
E4	2	1	BILL BARRETT CORP <BBG.N>: BMO CUTS PRICE TARGET TO 8 FROM <TRT>	Reference to other numerals
E5	3	2	Maersk Drilling wins \$ <TRT> mln contract from Eni	Main event
E6	7	6	OCC SAYS EXCHANGE-LISTED OPTIONS VOLUME REACHED <TRT> CONTRACTS IN MAY	Varying amounts

Table 5: Error analysis (T: truth; P: prediction; 0: the decimal; 7: magnitude greater than 6)

	0	1	2	3	4	5	6	7
Market Drilling	0	1	0	2	0	0	0	0
wins	0	489	177	266	17	1	3	1
mln contract	0	39	46	30	1	0	0	0
Eni	0	51	11	2	12	0	4	0

Table 6: Co-occurrence statistics of (E5)

Table 6 lists the co-occurrence statistics of the keywords and each class label for (E5). From the prediction of the 2nd magnitude, we find that models do not focus on the most frequent word (wins) but on the key term (mln contract) in this comment. Besides, the influence of company names (Maersk Drilling and Eni) may be less than that of the key term. Therefore, we infer that the models can capture the main event in a market comment.

In (E6), the <TRT> label should be replaced by 377,539,997. Volume patterns vary, however, for different financial instruments. For example, the trading volume of Alphabet Inc. (GOOG) was about 4,760K (the 7th magnitude) on 2018/04/24 but about 899K (the 6th magnitude) on 2018/05/25. This indicates that trading volume can be diverse even for the same stock.

The task setting in this paper is the coarse-grained setting for numeracy. More fine-grained settings toward numeracy can be extended in future works. For example, leveraging the taxonomy of the numeral information (Chen et al., 2018) and understanding the relationship between the named entities and the numbers (Chen et al., 2019) may be able to improve the performance of learning numeracy.

5 Discussion

Fake news has brought negative effects, especially in the 2016 U.S. presidential election (Bakir and McStay, 2018). In the financial domain, even one piece of negative information can cause a stock price to crash. If someone with bad intentions

introduces fake information about a company, its stock price can be influenced violently. Especially during trading hours, investors might not have enough time to verify such news, and the company could not declare its falsehood rapidly enough. In this section, we provide a first report of the simulated experimental results focusing on financial market comments, suggesting the capability of the models to detect such exaggerated numerals in market comments. We further experiment on The Examiner dataset¹ to show the numeracy of models toward the article titles of crowdsourced journalism.

5.1 Exaggerated Numeral Detection

To examine the BiGRU models reasoning ability, we multiply the numerals in market comments by different distortion factors. Then, the model aims to detect whether a numeral is correct, overstated or understated. For example, 138 in (S11) with 10% distortion factor will become 124.2 (-10%) and 151.8 (+10%), and both are considered as exaggerated numerals.

(S11) *SPLUNK INC <SPLK.O> SEES Q2 2016 REVENUE \$138 MLN TO \$140 MLN*

In this experiment, we release the boundary limitation, and test the numeracy for all real numbers. For instance, the altered results of 138 with 10% distortion factor are in the same magnitude, and that with 30% distortion factor, 96.6 and 179.4, are in different magnitude. Table 7 lists the experimental results. We find that the model obtained better performance for numerals distorted by more than 50%, with more confusion in the range below that. Furthermore, according to the micro- and macro-averaged F1 scores, the performance is similar among the three different cases (i.e., overstated, understated, and correct).

In summary, our experiments show that we can not only learn the concept of magnitude, but also

¹<https://www.kaggle.com/therohk/examine-the-examiner>

Distortion factor	Micro-F1	Macro-F1
$\pm 10\%$	58.54%	57.87%
$\pm 30\%$	56.94%	56.11%
$\pm 50\%$	57.69%	56.85%
$\pm 70\%$	70.92%	70.85%
$\pm 90\%$	76.91%	76.94%

Table 7: Results for exaggerated numeral detection.

M	0	1	2	3	4	5	6	7
%	0.08	35.18	30.94	8.71	24.21	0.57	0.31	0.01

Table 8: Distribution of numerals in the title dataset. M.: magnitude; 7: $M > 6$.

discover the concept of the reasonableness of the numerals in financial tweets. This kind of numeracy can be applied to many potential application scenarios, e.g., avoiding fat-finger error in the financial market, detecting the carelessly wrong of dosage in the doctor’s advice, and so on.

5.2 Numeracy in Open-Domain Article Titles

The distribution of the numerals in the article title dataset is shown in Table 8. Comparing with the distribution of market comments, few article titles use decimal. On the other hand, writers of articles use more 4th-magnitude numerals than those in market comments. Total 23.25% of titles contain at least one numeral. Although the proportion is lower than that in the financial narrative, it still shows that numerals are important and informative in the general description.

The experimental results are shown in Table 9. The BiGRU model outperforms the other models in both Micro-F1 and Macro-F1. Based on the experimental results on both datasets, BiGRU may be the best model for learning numeracy. In general, models perform relatively worse in the article title dataset than in the market comment dataset. The performance gaps may be caused by the following reasons. (1) The topics in titles are more diverse than those in market comments. (2) To attract more clicks, title writers may use a catchy numeral, which can be an exaggerated number. The illogical numbers may not only confuse humans, but also models. We leave the in-depth experiment on applying numeracy to detect illogical numbers in the future work, because more fine-grained annotations are needed.

We further adopt the BiGRU model to test the numeracy with the cross-source data, i.e., one

Model	Micro-F1	Macro-F1
LR	62.49%	30.81%
CNN	69.27%	35.96%
GRU	70.92%	38.43%
BiGRU	71.49%	39.94%
CRNN	69.50%	36.15%
CNN-capsule	63.11%	29.41%
GRU-capsule	70.73%	33.57%
BiGRU-capsule	71.49%	34.18%

Table 9: Experimental results of titles.

Training	Test set	Micro-F1	Macro-F1
Comment	Title	31.38%	11.08%
Title	Comment	25.59%	10.58%

Table 10: Results of learning cross-source numeracy.

serves as the training set, and the other as the test set. The results in Table 10 show the difficulty of transferring numeracy toward different sources.

6 Conclusion

We present a novel task of learning numeracy with the Numeracy-600K,² including the market comments and the article titles. The experimental results show that NN models can learn the proper range for a target numeral from contextual information. An experiment on an application scenario of exaggerated numeral detection suggests the capability of the proposed NN models. In future work, we plan to extend our work to further applications such as detecting exaggerated statements by investors in social media data.

Acknowledgments

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-107-2634-F-002-011-, MOST-108-2634-F-002-008-, and MOST 107-2218-E-009-050-, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05. This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

²<https://github.com/aistairc/Numeracy-600K>

References

- Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, 6(2):154–175.
- Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral attachment with auxiliary tasks. In *The 42nd International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM.
- Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143. IEEE.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.
- Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2017. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.
- Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. 2017. Learning to generate market comments from stock prices. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1374–1384.
- Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.
- Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115.
- Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Proceedings of the 2018 Web Conference*, pages 517–524.
- Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018a. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857.
- Yequan Wang, Aixin Sun, Jialong Han, Ying Liu, and Xiaoyan Zhu. 2018b. Sentiment analysis by capsules. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1165–1174.

A Appendices

We report the details for the replication of the experiments in the following appendices.

A.1 Convolutional Neural Network (CNN)

We construct a CNN model for numeracy. Modified from the CNN for sentence classification (Kim, 2014), in our model, each word in the input sentence is represented as a d -dimensional vector, and all the words are concatenated in as a $d \times l$ matrix, where l denotes the sentence length. The target numeral to be inferred is replaced with a special token <TRT>. The output of our CNN model is a softmax layer that generates the probability distribution over the magnitudes for the target numeral.

The details of our CNN model are described as follows. The size of the first layer, the embedding layer, is set as $d = 300$. We set $l = 73$, which is the longest sentence in the dataset. Padding is performed for shorter sentences. The second layer is a convolutional layer with filter size 8. The third layer is a fully connected layer with dimension 32, which functions as a max-pooling layer. To avoid overfitting, a dropout layer is added with a dropout rate of 0.3. Finally, two activation functions, the rectified linear unit (ReLU) and softmax, are used in the last two layers. We chose to use the Adam optimizer.

A.2 Gated Recurrent Unit (GRU)

We construct an RNN-based model for numeracy with GRU. The tokens in the sentence are input as a sequence. Each token is represented as a d -dimensional vector. The target numeral is replaced with the special token <TRT>. The architecture of the GRU model in this paper consists of a 300-dimensional embedding layer, a 64-dimensional

GRU layer, and a dropout layer with a dropout rate of 0.3. The final two layers and the optimizer are the same as those in the CNN model.

A.3 Bidirectional GRU (BiGRU)

The bidirectional RNN model, BiGRU, merges the outputs from both directions of the GRU model. Because units of measurement provide the important clues for numeral, a bidirectional architecture is expected to be useful with the right to left inputs. For example, the difference between (C1) and (C2) is the unit of measurement (i.e., *POINTS* and *PERCENT*), and it leads to different results of the magnitude of numerals.

(C1) *DOW JONES <.DJI> UP 8.70
POINTS*

(C2) *DOW JONES <.DJI> UP 0.05
PERCENT*

A.4 Convolutional Recurrent Neural Network (CRNN)

In our CRNN model, a CNN layer extracts features for each segment. Then, a max-pooling layer in the CNN model is replaced by an RNN layer and aggregates the extracted features. To examine whether replacing the pooling layer with the RNN layer can improve performance in our task, we keep the other components of the CRNN model the same as those in the CNN model, and replace the max-pooling layer with the 64-dimension BiGRU layer.

A.5 CNN-capsule

We also introduce one of the latest architectures, capsule network, to the task of numeracy. We combine the capsule network with either of the CNN and the GRU models. The structure of the CNN-capsule model begins with a 300-dimensional embedding layer. The second layer is a convolutional layer having a kernel size of 9 and using the ReLU activation function. The third layer, called the primary layer, is used to retain the order of context information, including one convolutional layer with 32 channels. Finally, the capsule layer outputs an $n \times dim$ matrix, where n is the number of classes, set to 8 for this paper, and dim is the dimension of each capsule, set to 16.

A.6 GRU-capsule

The GRU-capsule model begins with a 300-dimensional embedding layer, followed by a 64-

dimensional GRU layer, which returns the full sequence of outputs. To compare the impacts of the CNN and RNN frameworks in the CapsNet architecture, we keep the primary and capsule layers the same as those in the CNN-capsule model.

A.7 BiGRU-capsule

We further explore the bidirectional GRU model with the addition of capsule network. The BiGRU-capsule model consists of a 300-dimensional embedding layer, bidirectional GRU layers with a 64-dimensional hidden state, and the primary and capsule layers described above.