HYPHEN: Hyperbolic Hawkes Attention For Text Streams

Shivam Agarwal[*] **Ramit Sawhney**[*] **Sanchit Ahuja, Ritesh Soun, Sudheer Chava**
Financial Services Innovation Lab, Georgia Institute of Technology
rsawhney31@gatech.edu, sudheer.chava@scheller.gatech.edu

Abstract

Analyzing the temporal sequence of texts from sources such as social media, news, and parliamentary debates is a challenging problem as it exhibits time-varying scale-free properties and fine-grained timing irregularities. We propose a Hyperbolic Hawkes Attention Network (HYPHEN), which learns a data-driven hyperbolic space and models irregular powerlaw excitations using a hyperbolic Hawkes process. Through quantitative and exploratory experiments over financial NLP, suicide ideation detection, and political debate analysis we demonstrate HYPHEN's practical applicability for modeling online text sequences in a geometry agnostic manner.

1 Introduction

Text stream modeling is a critical problem that helps analyze trends over a variety of applications spanning finance (Oliveira et al., 2017), healthcare (Baytas et al., 2017), and political discourses (Sawhney et al., 2021c). However, analyzing such text sequences poses several challenges. First, modeling individual text items may not be informative enough since text sequences display a *sequential context dependency*, where analyzing them together in succession provides better contextual representation (Hu et al., 2018). Second, timing plays an essential role in online stream modeling as users quickly react to new information (Sawhney et al., 2021a). For instance, in stock markets, reacting a second slower than other investors can lead to massive losses (Scholtus et al., 2014). A fundamental limitation in existing RNN methods is that it ignores the natural fine-grained timing irregularities in streams (Foucault et al., 2016; Eysenck, 1968).

Social theories show that from a vast volume of texts in a stream, only a few are powerful enough to heavily influence the overall trend (Van Dijk, 1977; Gabaix, 2016). Such texts are rare and the excitation induced by them follows a powerlaw distribution which gives rise to scale-free properties (Zhao et al., 2010). For example, in political debates, there are a few rare highly-influential debates that heavily impact the overall voting decisions of citizens (Law, 2019). Further, the impact of such powerlaw excitations varies for each event. The presence of varying powerlaw dynamics from highly influential texts correlates with natural hierarchies and scale-free dynamics in text streams, making them difficult to model (Sala et al., 2018).

The good news is that hyperbolic learning has shown to better model such powerlaw dynamics compared to Euclidean learning over domains, including vision (Khrulkov et al., 2020) and NLP (Tifrea et al., 2019). However, existing works face two major limitations, 1) they ignore the timing irregularities in scale-free sequences and 2) they use a single hyperbolic space to encode varying levels of hyperbolic dynamics. Building on social theories, our contributions can be summarized as:

- We explore the hyperbolic properties of online streams and propose a Hyperbolic Hawkes Attention Network (HYPHEN) which jointly learns from the fine-grained timing irregularities and powerlaw dynamics of streams (§**2.2**).
- Building on social theories, HYPHEN learns the hyperbolic space based on the nature of the stream (§**2.1**). We introduce HYPHEN as a geometry agnostic model which can be applied on any downstream application.
- Through quantitative (§**4.1**) and exploratory (§**4.3**) experiments on four tasks spanning suicide ideation, political debate analysis, and financial forecasting over English and Chinese languages, we demonstrate the practical applicability of HYPHEN for stream modeling.[1]

[*] Equal contribution.

[1] We release HYPHEN's code at: https://github.com/gtfintechlab/HYPHEN-ACL

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 620 - 627
May 22-27, 2022 ©2022 Association for Computational Linguistics

2 Methodology

Problem Formulation: For a sequence of texts $[p_1 \dots, p_N]$ released at times $[t_1, \dots, t_N]$ sequentially, with $[t_1 < \dots < t_N]$, our target is to model this sequence in a time-sensitive fashion for a variety of downstream applications (§3).

2.1 Learnable Hyperbolic Geometry

Text sequences from social media and political discourses pose hierarchies (Sawhney et al., 2021a) i.e., the datasets represent a tree like structure which call for the use of hyperbolic spaces. Indeed, the volume of hyperbolic geometry grows exponentially, in contrast to Euclidean spaces where the growth is polynomial (Khrulkov et al., 2020), enabling hyperbolic spaces to capture the underlying scale-free properties of streams (Sala et al., 2018). However, text sequences exhibit a varying degree of scale-free dynamics, which a single geometry cannot capture (Gu et al., 2019). Thus, we seek to learn the optimal underlying geometry.

The hyperbolic space is a non-Euclidean space with a constant negative curvature c. To learn the optimal geometry, we aim to learn the curvature c, which controls the degree of hyperbolic properties represented by the space (Gu et al., 2019). Following (Ganea et al., 2018) we define the hyperbolic geometry with varying curvature c as $(\mathcal{B}, g_x^{\mathcal{B}})$, where the manifold $\mathcal{B} = \{x \in \mathbb{R}^n : c||x|| < 1\}$, is endowed with the Riemannian metric $g_x^{\mathcal{B}} = \lambda_x^2 g^E$, where the conformal factor $\lambda_x = \frac{2}{1-c||x||^2}$ and $g^E = \text{diag}[1, .., 1]$ is the Euclidean metric tensor. We denote the tangent space centered at point $\boldsymbol{x}$ as $\mathcal{T}_{\boldsymbol{x}}\mathcal{B}$. We generalize Euclidean operations to the hyperbolic space via Möbius operations.

Möbius Addition $\oplus$ for two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{B}$, is,

$$x \oplus y = \frac{(1 + 2c\langle x, y\rangle + c||y||^2)x + (1 - c||x||^2)y}{1 + 2c\langle x, y\rangle + c^2||x||^2||y||^2} \quad (1)$$

$\langle ., . \rangle$, $|| \cdot ||$ denotes the inner product and norm.

Exponential Map maps a tangent vector $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{B}$ to a point $\exp_{\boldsymbol{x}}(\boldsymbol{v})$ in the hyperbolic space,

$$\exp_{\boldsymbol{x}}(\boldsymbol{v}) = \boldsymbol{x} \oplus \left(\tanh\left(\frac{\sqrt{c}\lambda_{\boldsymbol{x}}||\boldsymbol{v}||}{2}\right)\frac{\boldsymbol{v}}{\sqrt{c}||\boldsymbol{v}||}\right) \quad (2)$$

Logarithmic Map maps a point $\boldsymbol{y} \in \mathcal{B}$ to a point $\log_{\boldsymbol{x}}(\boldsymbol{y})$ on the tangent space at $\boldsymbol{x}$,

$$\log_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{2}{\sqrt{c}\lambda_{\boldsymbol{x}}}\tanh^{-1}\left(\sqrt{c}||-\boldsymbol{x} \oplus \boldsymbol{y}||\right)\frac{-\boldsymbol{x} \oplus \boldsymbol{y}}{||-\boldsymbol{x} \oplus \boldsymbol{y}||} \quad (3)$$

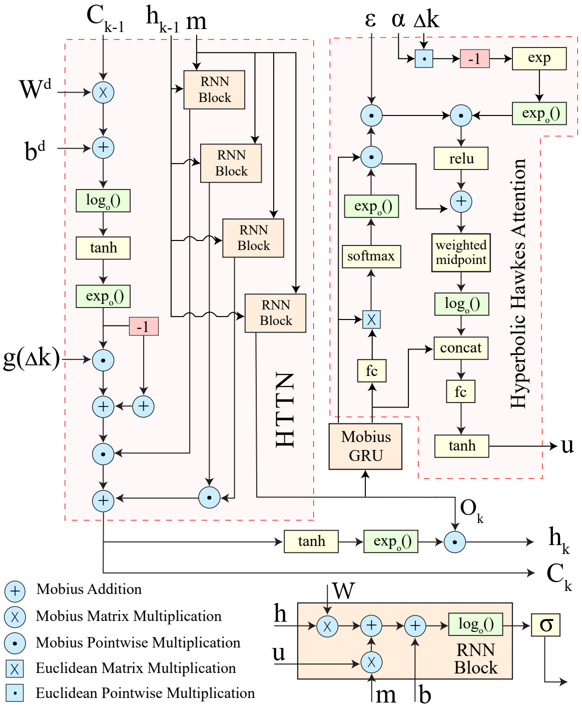

Figure 1: HYPHEN cell diagram and update rule.

Möbius Multiplication $\otimes$ multiplies features $\boldsymbol{x} \in \mathcal{B}^C$ with matrix $\boldsymbol{W} \in \mathbb{R}^{C' \times C}$, given by

$$\boldsymbol{W} \otimes \boldsymbol{x} = \exp_{\boldsymbol{o}}(\boldsymbol{W} \log_{\boldsymbol{o}}(\boldsymbol{x})) \quad (4)$$

Möbius Pointwise Product $\odot$ multiplies matrix $\boldsymbol{x} \in \mathcal{B}^C$ with matrix $\boldsymbol{y} \in \mathcal{B}^C$ pointwise,

$$\boldsymbol{x} \odot \boldsymbol{y} = \frac{1}{\sqrt{c}}\tanh\left(\frac{||xy||}{y}\text{arctan}^{-1}(\sqrt{c}||y||)\right)\frac{||xy||}{||y||} \quad (5)$$

2.2 HYPHEN: Hyperbolic Hawkes Network

Text Embedding Layer We use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to encode each text p_i to features $\hat{\boldsymbol{m}}_i = \text{BERT}(p_i) \in \mathbb{R}^d$ where $d = 768$, obtained by averaging the token level outputs from the final layer of BERT. To apply hyperbolic operations over text features $\hat{\boldsymbol{m}}_i$, we project it to the hyperbolic space via the exponential mapping $\exp_o(\cdot)$ given by, $\boldsymbol{m}_i = \exp_o(\hat{\boldsymbol{m}}_i)$

Hyperbolic Time Aware Temporal Network To encode the varying scale-free characteristics of text sequences, we introduce LSTMs over learnable hyperbolic spaces by leveraging Möbius operations (§2.1). Further, capturing fine-grained timing irregularities in text streams plays a crucial role for stream state modeling. For instance, the time interval between two debates can vary widely, from a few days to many months in parliamentary debates. Consequently, the ideologies and thought process of the speaker may change over time, reflecting a decay or increase in dependence on the speaker's previous speeches (Van Dijk, 2002).

To capture these time dependent intricacies in a learnable hyperbolic space, we modify the hyperbolic LSTM (Shimizu et al., 2021) as shown in Figure 1 into a hyperbolic time-aware temporal network (HTTN$(\cdot)$). Intuitively, the greater the time elapsed between text releases, the lesser the impact they should have on each other. Thus, for a given day k, HTTN applies a decaying function over Δk, the elapsed time between two texts $[p_k, p_{k-1}]$, transforming the time differences into weights:

$$\begin{aligned}
\boldsymbol{C}^s_{k-1} &= \exp_{\boldsymbol{o}}(\tanh(\log_{\boldsymbol{o}}(\boldsymbol{W}^d \otimes \boldsymbol{C}_{k-1} \oplus \boldsymbol{b}^d))) \\
\hat{\boldsymbol{C}}^s_{k-1} &= \boldsymbol{C}^s_{k-1} \odot g(\Delta k) \quad \text{Discounted short-term memory} \\
\boldsymbol{C}^T_{k-1} &= -\boldsymbol{C}^s_{k-1} \oplus \boldsymbol{C}_{k-1} \qquad \text{Long term memory} \\
\boldsymbol{C}^*_{k-1} &= \boldsymbol{C}^T_{k-1} \oplus \hat{\boldsymbol{C}}^s_{k-1} \qquad \text{Adjusted previous memory}
\end{aligned}$$

where $\boldsymbol{C}^s_{k-1}$ is the previous cell memory, $\boldsymbol{W}^d; \boldsymbol{b}^d$ are the network parameters, and $g(\cdot)$ is a heuristic decaying function. Following (Baytas et al., 2017) we set $g(\Delta k) = 1/\Delta k$. Using the adjusted previous memory $\boldsymbol{C}^*_{k-1}$, we define the current hidden state and current memory states for HTTN, with hyperbolic features m as:

$$\begin{aligned}
\widetilde{\boldsymbol{c}}_{\boldsymbol{k}} &= \sigma \log_{\boldsymbol{o}}(\boldsymbol{W}^c \otimes \boldsymbol{h}_{k-1} \oplus \boldsymbol{U}^c \otimes \boldsymbol{m}_k \oplus \boldsymbol{b}^c) \\
\boldsymbol{C}_k &= \boldsymbol{i}_k \odot \widetilde{\boldsymbol{c}}_{\boldsymbol{k}} \oplus \boldsymbol{f}_k \odot \boldsymbol{C}^*_{k-1} \qquad \text{(Current memory)} \\
\boldsymbol{h}_k &= \boldsymbol{o}_k \odot \exp_{\boldsymbol{o}}(\tanh(\boldsymbol{C}_k)) \qquad \text{(Current hidden state)}
\end{aligned}$$

where $\boldsymbol{W}^c; \boldsymbol{U}^c; \boldsymbol{b}^c$ are the learnable parameters, $\boldsymbol{i}_k; \boldsymbol{f}_k; \boldsymbol{o}_k$ are input, forget and output gates. Finally, given texts $[p_1, \ldots p_T]$ over a lookback period T, we define the update rule of HTTN as,

$$\boldsymbol{h_j} = \text{HTTN}(\boldsymbol{m_j}, \Delta j, \boldsymbol{h_{j-1}}); \quad j \in [1, T] \qquad (6)$$

where, h_j represents the hidden states of HTTN.

Hyperbolic Hawkes Attention Studies show that not all historical texts are equally informative and pose a *diverse influence* over the predictions (Sawhney et al., 2021c). We use a temporal hyperbolic attention mechanism (Luong et al., 2015) to emphasize texts likely to have a substantial influence. This mechanism learns attention weights β_i for each hiddden state $\boldsymbol{h_i} \in \overline{\boldsymbol{h}} = [\boldsymbol{h_1}, \ldots, \boldsymbol{h_T}]$ as,

$$\beta_j = \text{Softmax}\Big(\exp\left(\log_{\boldsymbol{o}}(\boldsymbol{h_j})^{\mathrm{T}}(\boldsymbol{W} \log_{\boldsymbol{o}}(\overline{\boldsymbol{h}}))\right)\Big) \quad (7)$$

where, $\boldsymbol{W}$ denotes learnable weights.

Next, we enhance the temporal hyperbolic attention using the Hawkes process (Mei and Eisner, 2017) and propose a hyperbolic Hawkes attention mechanism. The Hawkes process is a temporal point process that models a sequence of arrival of texts over time. Each text item *"excites"* the process in the sense that the chance of a subsequent arrival is increased for some time. Studies (Zuo et al., 2020; Sawhney et al., 2021b) show that the Hawkes process can be used to model text sequences from social media and discourses. The hyperbolic Hawkes attention mechanism learns an excitation parameter ϵ corresponding to excitation induced by text p_j and a decay parameter α to learn the decay rate of this induced excitement. Formally, we use an Einstein midpoint (Ungar, 2005) to aggregate hidden states $\overline{h}$ via Hawkes process as,

$$\boldsymbol{u} = \texttt{HYPHEN}(\{p_i, t_i\}_{i=1}^T) = \sum_j \frac{\beta_j \gamma(\boldsymbol{q}_j)}{\sum_\tau \beta_\tau \gamma(\boldsymbol{q}_\tau)} \boldsymbol{q}_j \qquad (8)$$

$$\boldsymbol{q_j} = \beta_j \odot \boldsymbol{h_j} \oplus \epsilon \odot \exp_o\left(\text{ReLU}(\log_o(\boldsymbol{h_j}))\right) \odot e^{-\alpha \Delta k} \qquad (9)$$

where, $\gamma(\boldsymbol{q}_j) = \frac{1}{\sqrt{1 - ||\boldsymbol{q}_j||^2}}$ are the lorentz factors.

3 Applications and Tasks

Political Stance Prediction Parliamentary debates consist of responses from politicians over a motion. Following (Sawhney et al., 2020), we aim to classify the stance of a speaker as 'Aye'/'No' on a motion based on their historic speeches. We evaluate on the ParlVote dataset (Abercrombie and Batista-Navarro, 2020) comprising of 33,461 UK debate transcripts of 1,346 politicians.

Financial NLP We aim to predict future stock trends based on the historic texts about a stock. Following (Sawhney et al., 2021a) we regress the future volatility of a stock defined as $\lambda = ln(|\frac{p_i - p_{i-1}}{p_{i-1}}|)$, where p_i is the closing price. We evaluate on the S&P (Xu and Cohen, 2018) containing 88 stocks with 109,915 tweets and the China Stock Exchange (CSE) (Huang et al., 2018) containing 90,361 Chinese news articles for 85 stocks.

Suicide Ideation Following (Sawhney et al., 2021d), we aim to detect suicidal intent in a tweet given historic tweets from a user. We use the data from (Mishra et al., 2019) containing 32,558 user timelines and 2.3M texts.

Table 1: Performance comparison with baselines (mean of 40 runs). * indicates improvement over SOTA is significant ($p<0.01$) under Wilcoxon's signed rank test.

Model	PVote MCC ↑	SI MCC ↑	CSE MSE ↓	S&P MSE ↓
MLP(2018)	0.36	0.24	2.91	0.38
LSTM(1997)	0.52	0.28	2.88	0.34
HAN(2019)	0.50	0.29	2.85	0.31
H-LSTM(2020)	0.53	0.29	2.87	0.33
FAST(2021e)	0.51	0.30	2.86	0.32
HT-LSTM(2021a)	0.55	0.31	2.68	0.31
HYPHEN (Ours)	**0.63***	**0.44***	**2.68**	**0.29***

Table 2: Ablation study over HYPHEN (mean of 40 runs). *,†indicate improvement over HYPHEN-constant curvature and Euclidean (EUC) counterparts are significant ($p < 0.01$) under Wilcoxon's signed rank test.

Ablation Components	PVote MCC↑	SI MCC↑	CSE MSE↓	S&P MSE↓
LSTM	0.52	0.28	2.88	0.34
EUC-Time LSTM+Attn	0.51	0.30	2.86	0.32
EUC-Time LSTM+Hwks	0.54	0.33	2.83	0.32
HYP-time LSTM + Attn	0.58†	0.31†	2.73†	0.31†
HYPHEN-constant curvature	0.61†	0.36†	2.72†	0.30†
HYPHEN (Ours)	**0.63***†	**0.44***†	**2.68***†	**0.29***†

4 Results

4.1 Performance Comparison

We compare the performance of HYPHEN over financial, political, and healthcare tasks spanning English and Chinese languages in Table 1. We observe that HYPHEN generally outperforms most baseline methods by 10% on average. Overall, we note that methods that capture fine-grained timing irregularities in text sequences perform better (HYPHEN, FAST, HT-LSTM), validating our premise of using time-aware modeling. We postulate that HYPHEN's superior performance is due to, 1) learnable hyperbolic geometry and 2) time-aware hyperbolic Hawkes process. First, HYPHEN better encodes the varying hyperbolic properties of text sequences by learning a suitable data-driven curvature in contrast to other hyperbolic models (HT-LSTM), which constrain all sequences to a fixed hyperbolic space. Second, through hyperbolic time aware learning and Hawkes attention, HYPHEN better captures timing irregularities between the subsequent release of texts (Sawhney et al., 2021a). These observations collectively show the practical applicability and generalizability of HYPHEN for stream modeling.

4.2 Ablation Study

We contextualize the impact of various components of HYPHEN in Table 2. We note that augmenting RNN-based methods with attention leads to significant improvements ($p < 0.01$), as HYPHEN can better distinguish noise inducing text from relevant information (Sawhney et al., 2021e). Next, we observe significant ($p < 0.01$) improvements on using hyperbolic spaces to represent text streams, suggesting that the hyperbolic space better models the innate power-law dynamics and hierarchies in online text streams (Sala et al., 2018). Further, enriching the temporal attention with the Hawkes process leads to performance boosts, potentially because the Hawkes process better captures the excitation induced by influential texts. Finally, learning the underlying hyperbolic geometry benefits HYPHEN, allowing it to generalize to a variety of text streams with different hyperbolic properties.

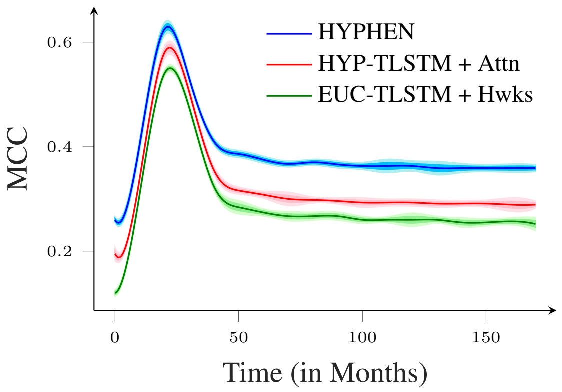

Figure 2: Sensitivity of HYPHEN to the lookback period T on political speaker state modeling.

4.3 Impact of Historical Context

We study the variation in HYPHEN's performance on political speaker state modeling corresponding to varying amounts of lookback periods T in Figure 2. First, without encoding the historic context, we observe that all models perform poorly. As we increase the lookback period, we note that Hawkes attention improves temporal attention, potentially because the Hawkes process decays the impact of very old texts enabling HYPHEN to focus on more recent debates which better reflects a speaker's temporal state. Further, with very large lookback periods, we observe a performance drop, likely because large amounts of context allow the inclusion of speeches from very old (stale) debates, which may not contribute significantly to the speaker's present state (Cullen et al., 2018). However, through hyperbolic Hawkes attention HYPHEN is able to filter out more crucial debates to an extent. In general, HYPHEN provides the best results with debates around ten months in the past (mid-sized lookbacks).

5 Conclusion

We explore the scale-free dynamics and timing irregularities of text streams. We propose HYPHEN which uses hyperbolic Hawkes attention and learns data-driven geometries to represent varying hyperbolic properties of streams. Through experiments on political, financial NLP, and healthcare tasks, we show the applicability of HYPHEN on 4 datasets.

Acknowledgements

We would like to thank the Financial Services Innovation Lab at Georgia Institute of Technology for their generous support.

6 Ethical Considerations

The sensitive nature of this work calls for careful deliberation of the risks and ethical challenges involved. While we only use publicly available user data, we emphasize the importance of preserving the privacy of the users involved (De Choudhury et al., 2016). We acknowledge that the predictive power of HYPHEN depends on the data, which is in tension with user privacy concerns. We carefully adopt the measures followed by Chancellor et al. (2016). Specifically, we operate within the acceptable privacy bounds (Chancellor et al., 2019) and considerations (Fiesler and Proferes, 2018) in order to avoid coercion and harmful interventions (Chancellor et al., 2019). We paraphrase and anonymize all samples in the suicide ideation detection detection dataset using the moderate disguise scheme (Bruckman, 2002; Fiesler and Proferes, 2018). We also perform automatic de-identification using named entity recognition to identify and mask personally identifiable information.

While one of our work's application is to aid in the early detection of suicidal users and early intervention, it is imperative that any interventions be well-thought, failing which may lead to counter-helpful outcomes, such as users moving to fringe platforms, which would make it harder to provide assistance (Kumar et al., 2015). Care should be taken so as not to create stigma, and interventions must be carefully planned by consulting relevant stakeholders such as clinicians, designers, and researchers (Chancellor et al., 2016), to maintain social media as a safe space for individuals looking to express themselves (Chancellor et al., 2019).

References

Gavin Abercrombie and Riza Batista-Navarro. 2018. 'aye' or 'no'? speech-level sentiment analysis of hansard uk parliamentary debate transcripts.

Gavin Abercrombie and Riza Batista-Navarro. 2020. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 65–74, New York, NY, USA. Association for Computing Machinery.

Gary Bécigneul and Octavian-Eugen Ganea. 2018. Riemannian adaptive optimization methods. *CoRR*, abs/1810.00760.

Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4(3).

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*.

Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Ailbhe Cullen, Andrew Hines, and Naomi Harte. 2018. Perception and prediction of speaker appeal – a single speaker study. *Computer Speech & Language*, 52:23 – 40.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hans J Eysenck. 1968. *The psychology of politics*, volume 2. Transaction publishers.

Casey Fiesler and Nicholas Proferes. 2018. “participant” perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.

Thierry Foucault, Johan Hombert, and Ioanid Roşu. 2016. News trading and speed. *The Journal of Finance*, 71(1):335–382.

Xavier Gabaix. 2016. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30(1):185–206.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. In *NeurIPS*, pages 5350–5360.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. 2019. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. 2019. Hyperbolic attention networks. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM ’18, page 261–269, New York, NY, USA. Association for Computing Machinery.

Jieyun Huang, Yunjia Zhang, Jialai Zhang, and Xi Zhang. 2018. A tensor-based sub-mode coordinate algorithm for stock prediction.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. 2020. Hyperbolic image embeddings. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext & Social Media*, pages 85–94.

Tara Law. 2019. The most important presidential debates in american history, according to historians. `https://time.com/5607429/most-important-debates/`. Accessed: 2021-09-15.

Federico López and Michael Strube. 2020. A fully hyperbolic neural model for hierarchical multi-class classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 460–475, Online. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Hongyuan Mei and Jason Eisner. 2017. The neural hawkes process: A neurally self-modulating multivariate point process.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2017. The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73:125–144.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. 2018. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4457–4466. PMLR.

Ramit Sawhney, Shivam Agarwal, Megh Thakkar, Arnav Wadhwa, and Rajiv Ratn Shah. 2021a. *Hyperbolic Online Time Stream Modeling*, page 1682–1686. Association for Computing Machinery, New York, NY, USA.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, Tyler Derr, and Rajiv Ratn Shah. 2021b. Stock selection via spatiotemporal hypergraph attention network: A learning to rank approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):497–504.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Shah. 2021c. Tec: A time evolving contextual graph model for speaker state analysis in political debates. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3552–3558. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021d. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190, Online. Association for Computational Linguistics.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2020. GPolS: A contextual graph-based language model for analyzing parliamentary debates and political cohesion. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4847–4859, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021e. FAST: Financial news and tweet based time aware network for stock trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2164–2175, Online. Association for Computational Linguistics.

Martin Scholtus, Dick van Dijk, and Bart Frijns. 2014. Speed, algorithmic trading, and market quality around macroeconomic news announcements. *Journal of Banking and Finance*, 38:89 – 105.

Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. 2021. Hyperbolic neural networks++. In *International Conference on Learning Representations*.

Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. 2019. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*.

Abraham A Ungar. 2005. *Analytic hyperbolic geometry: Mathematical foundations and applications*. World Scientific.

Teun A Van Dijk. 2002. Political discourse and political cognition. *Politics as text and talk: Analytic approaches to political discourse*, 203:203–237.

Teun Adrianus Van Dijk. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. Longman London.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

Xiaojun Zhao, Pengjian Shang, and Yulei Pang. 2010. Power law and stretched exponential effects of extreme events in chinese stock markets. *Fluctuation and Noise Letters*.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer hawkes process. In *International Conference on Machine Learning*, pages 11692–11702. PMLR.

A Experimental Setup

A.1 Datasets

- **US S&P (Xu and Cohen, 2018):** US S&P stocks are categorized into 9 industries: basic materials, consumer goods, healthcare, services, utilities, conglomerates, financial, industrial goods and technology. US S&P dataset contains text data and historical prices of 88 stocks which includes all 8 stocks in conglomerates and the top 10 stocks by market capitalization in each of the other industries. The text data comprises tweets from 01/01/2014 to 01/01/2016. Following (Xu and Cohen, 2018) we split the US S&P temporally based on date ranges from 01/01/2014 to 01/08/2015 for training, 01/08/2015 to 01/10/2015 for validation, and 01/10/2015 to 01/01/2016 for test.

- **China and Hong Kong (CSE) (Huang et al., 2018):** China and Hong Kong (CSE) dataset consists of news headlines of 85 top-traded stocks listed on the Shanghai, Shenzhen, and Hong Kong Stock Exchange from January 2015 to December 2015. The qualitative data comprises of 90,361 Chinese financial news headlines. We split the China & HK dataset temporally based on date ranges from 01/01/2015 to 31/08/2015 for training, 01/09/2015 to 30/09/2015 for validation, and 01/10/2015 to 01/01/2016 for testing all models.

- **ParlVote (Abercrombie and Batista-Navarro, 2020):** Following (Sawhney et al., 2020) we evaluate political stance detection on the ParlVote dataset. This record consists of debate transcripts from the UK House of Commons obtained under an open Parliament license. Following (Abercrombie and Batista-Navarro, 2020) we remove non-speech elements from the transcripts and the original casing is preserved. ParlVote consists of 33,461 transcripts from May 7th 1997 to November 5th 2019. The average number of tokens in a ParlVote speech is 760.2 ± 901.3. Based on a speaker's vote to their speech, transcripts are labeled as 'Aye' and 'No' representing positive and negative stance respectively. The dataset is fairly balanced, consisting of 53.57% ‘Aye’ and 46.43% ‘No’ labels. We split the dataset temporally to obtain 70%, 15% and 15% of the data for training, validation and testing respectively.

- **Suicide Ideation.** (**Sawhney et al., 2021d**): The Suicide ideation dataset is built upon the existing Twitter tweets database of (Mishra et al., 2019). The dataset consists of tweets of 32,558 unique users, spanning over ten years of historical tweets from 2009 to 2019. Out of all the tweets, 34,306 tweets were identified as having potential suicide ideation words. These tweets were then manually annotated by two psychologists under the supervision of a head psychologist and 3984 tweets were actually identified as having suicidal tendencies. The same preprocessing techniques were employed on the dataset as done by Sawhney et al. (2021d).

A.2 Evaluation Metrics

Matthews correlation coefficient: The Matthews correlation coefficient (MCC) produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset. We use MCC to evaluate on suicide ideation detection and political speech classification.

Mean squared error: To evaluate the volatility regression performance, we adopt the Mean Squared Error (MSE) to compute the error between actual and the predicted volatility values.

A.3 Baseline Models

We compare HYPHEN with the following baselines:

- **MLP**: A Bag of Words model that uses unigram textual features as input along with the TF-IDF vectors which are fed into a multi-layer perceptron (Abercrombie and Batista-Navarro, 2020).

- **LSTM** : An RNN architecture capable of learning long term sequential dependencies (Hochreiter and Schmidhuber, 1997).

- **HAN**: Transformer model with hyperbolic activations and attention which utilises hyperbolic geometry for both computation and aggregation of attention weights (Gulcehre et al., 2019).

- **H-LSTM**: A RNN based model for sequential data with an attention mechanism operating in the hyperbolic space (López and Strube, 2020).

- **FAST**: A time-aware LSTM network capable of modeling the fine grained temporal irregularities in textual data (Sawhney et al., 2021e).

- **HT-LSTM**: Hierarchical Time-aware hyperbolic LSTM network leverages the hyperbolic space for encoding scale-free nature of a text stream (Sawhney et al., 2021a).

A.4 Training Setup

We have performed all our experiments on Tesla GPU. We performed a grid search for all our models and selected the best values based on the validation MCC/MSE. We followed the same preprocessing techniques as suggested by the dataset authors. We explored the lookback window length $T \in [2, 20]$ and the hidden state dimensions in $\in (64, 128, 256)$. We grid searched our learning rates in $\in (1e-5, 5e-4, 1e-3)$. We used Riemannian Adam (Bécigneul and Ganea, 2018) as our optimizer.