EuroVis Workshop on Visual Analytics (2020)
K. Vrotsou and C. Turkay (Editors)

Interactive Visualization of AI-based Speech Recognition Texts

Tsung Heng Wu and Ye Zhao[†] and Md Amiruzzaman

Kent State University

Abstract

Speech recognition technology has achieved impressive success recently with AI techniques of deep learning networks. Speech-to-text tools are becoming prevalent in many social applications such as field surveys. However, the speech transcription results are far from perfection for direct use in these applications by domain scientists and practitioners, which prevents the users from fully leveraging the AI tools. In this paper, we show interactive visualization can play important roles in post-AI understanding, editing, and analysis of speech recognition results by presenting specified task characterization and case examples.

1. Introduction

Speech recognition is an important field in computational linguistics [CRS05, CFL13]. For many years, researchers have developed a variety of technologies and tools to identify words and phrases in spoken language [JM14, BMG*16, HM15a]. Recently, AI techniques, especially deep learning networks, have become revolutionary as they outperform previous methods and lead to high quality and low error rate in the speech to text results [HDY*12, MLJ*14]. Cloud-based speech to text services has been provided by many big companies such as Microsoft [Mic], Google [Goo19], etc. using deep learning models. Users from multiple domains are eager to utilize these AI tools for real-world applications such as conducting field surveys and collecting user opinions [BZK12, HM15b, Muh15]. However, the transcription results still have a set of practical problems including: (1) A full speech is recognized as a group of fragments, which usually do not represent natural sentences or paragraphs from the speaker; (2) Errors of audio recognition is inevitable and the quality varies greatly; (3) The confidence scores of words and fragments given by the speech recognition algorithms sometimes do not reflect the real probability of misrecognition. These problems have already hindered the more widespread use of speech to text tools [KRS17]. Domain scientists face challenges to effectively complete the following tasks in collecting lengthy audios from multiple speakers:

- Understanding the characteristics of speeches as well as speakers;
- Manually editing the massive speech-to-text results;
- Analyzing and comparing multiple speeches and speakers.

Interactive visual exploration can help the users alleviate the tasks in post-AI processing [KAKC17, KCK*18]. Fig. 1 illustrates the pathway of data processing, where a visual exploration loop is shown in red. It can enhance the usability of AI-based speech recognition tools. This short paper contributes to this path in two facets:

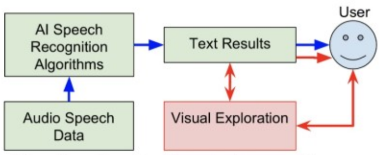


Figure 1: *Visual exploration loop in the speech-to-text pathway.*

- We identify the characteristics and shortcomings of AI-based speech-to-text results. Then, we propose major directions that visualization techniques can promote their usability.
- We develop a prototype to showcase the usefulness of visualization tools, which combines visual metaphors with semantic and sentimental analysis.

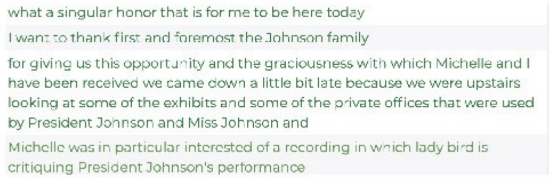
The purpose of this paper is to show how visual analytics can help to address the uncertainty issues in AI-based speech recognition.

2. Related Work

Traditional speech recognition systems use hidden Markov and Gaussian mixture models to represent the acoustic input in speech recognition [JM14]. Well trained deep neural networks with many hidden layers, possibly in combination with hidden Markov models, have outperformed on speech recognition benchmarks [GMH13]. Recurrent neural networks (RNNs) model speech as a dynamic time process, whose hidden state is a function of all previous hidden states [HDY*12]. Many research and commercial products, such as Microsoft, Google, Apple, IBM, and TensorFlow, have become usable for acoustic modeling and speech recognition [HDY*12].

[†] Corresponding author: zhao@cs.kent.edu

DOI: 10.2312/eurova.20201091



what a singular honor that is for me to be here today

I want to thank first and foremost the Johnson family

for giving us this opportunity and the graciousness with which Michelle and I have been received we came down a little bit late because we were upstairs looking at some of the exhibits and some of the private offices that were used by President Johnson and Miss Johnson and

Michelle was in particular interested of a recording in which lady bird is critiquing President Johnson's performance

Figure 2: *Speech-to-Text result of google cloud tool. Multiple fragments (separated by background color) have different lengths and confidence values.*

Visualization of neural networks has been an emerging topic for inspecting the training, improving the models and understanding the results ([RFFT17, LSL*17, ZZ18, HKPC18]). While most techniques work on vision datasets with a focus on CNNs (convolutional neural networks), RNN visualization tools are developed for linguistic, biological, and vision tasks. They focus on analyzing hidden state properties, phrase structure, and chord progressions, and so on [KJFF15, SGPR18, WPW*11].

Visualization can be useful to temporal event sequence data visualization, specifically to identify potential privacy issues in event-based or time-varying data [CWM19]. Moreover, researchers have presented various approaches in visual analytics of text corpora [LWC*18]. Visual sentiment analysis has been addressed where word clouds and other techniques are employed [KK15]. However, visualization techniques have not been applied to speech-to-text recognition results. This paper identifies the unique features and requirements of the AI outputs and visualizes them in a few examples which can help end-users better utilize these AI tools.

3. AI-based Speech to Text Outcome and Attributes

3.1. Text Fragments and Confidence

AI tools transcribe an audio speech to a text document (i.e., transcript) consisting of a list of *speech fragments*. Each fragment is a natural language segment of the speaker's narration based on their talking speed, stop, and other attributes. It consists of multiple terms (i.e. keywords) while each term has a unique term speaking length (**audio length**), which is different from its text word length. As shown in the example in Fig. 2, these fragments do not necessarily represent sentences as in written documents. The list is not easy to read and understand in comparison to a written document.

Second, the AI model usually provides a **term confidence** score for each term. Moreover, a **fragment confidence** score is also given for one speech fragment. These confidence scores assess the reliability of automatic speech transcriptions. They provide cues to the audio recognition errors for applications [RLGW18]. These confidence information and related errors need to be addressed with *human in the loop*. Moreover, each term may be marked with a **speaker tag** which indicates the AI recognized different speakers.

An important feature is that recognition errors may not be well indicated by the confidence scores. Other features such as audio length need to be used to help users address the errors, for example, a very long-recognized term usually refers to a failed word recognition.

3.2. Semantic and Sentiment Attributes

The attributes of term/fragment length and confidence can be visualized together with semantic and sentiment information, so as to promote deep understanding and allow quick revision. We have computed the following attributes (to be extended by using more NLP and text mining tools):

- Term frequency: The keywords in the text are ranked by their appearance frequency to identify top term in a speech;
- Term and fragment sentiment: Sentiment analysis can identify whether the expressed opinion in a document or a sentence is positive, negative, or neutral (e.g., AFINN [Nie11], MAN [JTL*20], SentiDiff [WNY19]). This is very helpful to discover speakers' attitudes and emotions. In this paper, we apply a sentiment analysis tool to discover the sentiment score of each fragment and each keyword [NPM]. AFINN English word list is utilized where each term is rated as an integer between minus five (negative) and plus five (positive) [Nie11]. Each fragment is given a sentiment score by summing up the sentiment integers of all the terms.
- Term Entropies: One same term in a speech can appear multiple times (N). Each time it may have different attributes ($d_i, i \in (1..N)$), such as audio length and confidence. The entropy of one term attribute shows the diversity of this attribute throughout the speech. For example, a high term entropy of audio length indicates that a speaker uses diverse vocal lengths for the same word. We compute the entropies of different attributes as $H = -\sum_{i\in(1..N)} p(d_i)\log p(d_i)$, where $p(d_i) = \frac{d_i}{\sum_{i\in(1..N)}(d_i)}$.

4. Identification of Visualization Tasks and Functions

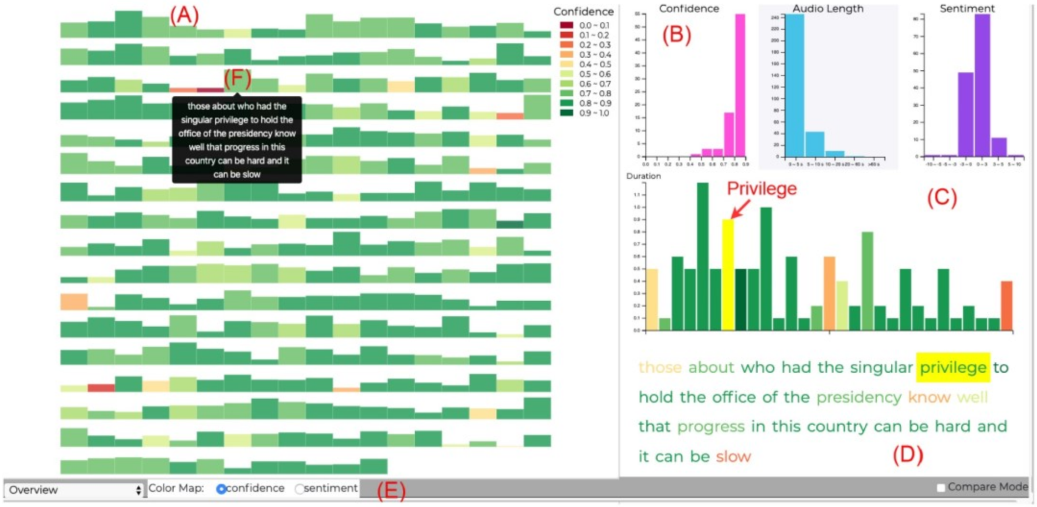
We talked with three social scientists (in public health, crime, urban study, and disaster management) who have used speech recognition in their work. From the requirement analysis, several tasks are identified for interactive visualization systems including (i.e., broken down into *speech fragments*):

- T1: Helping users quickly understand the structures and characteristics of the recognized text;
- T2: Allowing users to discover and compare semantics of speeches and sentiments of speakers;
- T3: Guiding the revision of the transcript to achieve better quality for downstream applications.

Addressing each task of T1 to T3 by visualization techniques and systems is challenging - e.g., for T3, developing an interactive interface for users to quickly and effectively edit and revise a large number of transcripts requires intensive system design, development, and evaluation. In this short paper, we coordinate the following visualization functions in an integrated prototype:

Transcript and Fragment Visualization: The speech-to-text results are visualized based on the recognized speech fragments with multiple attributes. Users can study the structure and patterns, and then drill-down to details by listening to the audio clip and visualizing the terms in the fragments.

Term Visualization: The keywords are visualized to show and integrate important attributes such as frequency, audio length, and con-



those about who had the singular privilege to hold the office of the presidency know well that progress in this country can be hard and it can be slow

those about who had the singular privilege to hold the office of the presidency know well that progress in this country can be hard and it can be slow

Figure 3: *Visual interface with Obama's Civil Right Speech 2015. (A) A bar page showing the confidence values of fragments in this speech (i.e., green color means high confidence and red color means low confidence). Hovering over one fragment f shows the text; (B) Histogram charts of speech attributes; (C) The word bars of f: each bar's height shows the word's audio length and color shows the word confidence. The word "privilege" has low confidence; (D) The text view of (F) "privilege" is highlighted (i.e., yellow color text and yellow color bar) while listening to the audio. (E) Control panel of visualization functions.*

fidence. Users can drill-down to study and compare related audio clips and their details.

Sentiment Visualization: Speaker's sentiments in a speech are visualized, together with audio features, so that users can identify the speaker's attitude and find important parts of the speech.

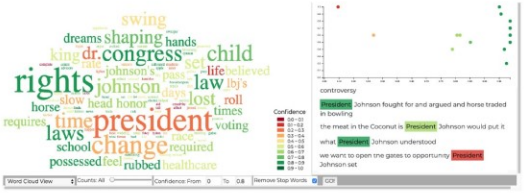
Comparative Visualization: Multiple speeches from the same or different speakers can be compared for insight discovery.

These functions are not complete for visual analytics tasks T1-T3, but instead are shown as a modest spur to encourage more valuable researches.

5. Prototype Visualization Design

The text results from AI recognition are visualized according to the following design requirements: (1) The whole picture of transcription structure should be displayed, with respect to the fragments and keywords; (2) The confidence and other attributes of fragments and keywords should be easily discerned; (3) The screen space should be well utilized to show these information.

Transcript Overview: A bar chart view is designed for the overview of a speech as shown in Fig. 3(A). It maps each fragment to a bar, and all the bars of one transcribed text are sequentially visualized, in order to represent as many data items as possible on the screen at the same time. The bar is colored and highlighted by different attributes selected by users. This view allows users to easily "browse" the document where hey can hover and click the bars to hear the raw audio and read the text. Instead of combining full text with confidence, the bar-chart visualization can present the full document in limited real estate. Fig. 3(A) visualizes the 2015 speech of President Obama about Civil Rights. The bar length is mapped to the audio length of a fragment, its color represents the fragment confidence. Meanwhile, a set of histograms of fragment confidence, audio length, and sentiments are shown in Fig. 3(B). Users can click on the histogram to highlight specific fragments of interests.



controversy

President Johnson fought for and argued and horse traded in bowling

the meat in the Coconut is President Johnson would put it

what President Johnson understood

we want to open the gates to opportunity President Johnson set

Figure 4: *Visualizing words in Obama's speech with their various attributes. Word size is the entropy of its audio lengths in multiple appearances and color is the minimum confidence (i.e., green color means high confidence and red color means low confidence).*

Fragment Drill-down: Users can click a fragment bar to study its details, whose audio clip is played. In Fig 3(A), a fragment f is clicked so that a fragment bar view is shown in Fig. 3(C). Here each bar refers to one word in this fragment. The words are visualized in Fig. 3(D). The terms are highlighted in yellow dynamically while the audio is playing. For example, the word "privilege" has low confidence and users can click it to listen to the raw audio repeatedly. This view allows users to identify erroneous recognition and correct them as needed.

Term visualization: The attributes of transcribed terms (keywords) are visualized to (1) find suspicious recognition; and (2) discover important semantic information with audio attributes. The visual design thus should give intuitive and fast cues for users to extract the information. A word cloud view is utilized to visualize the key-

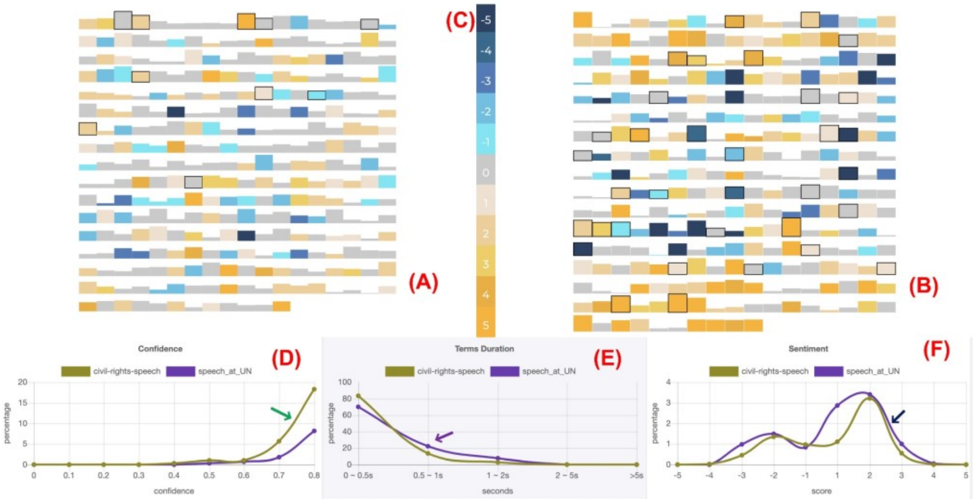
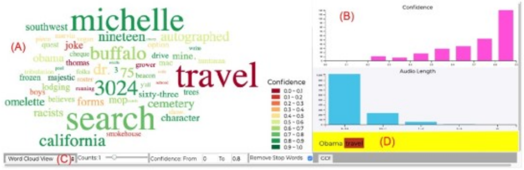


Figure 5: *Compare two speeches: President Obama on Civil Right 2015 Vs. President Trump in United Nations 2018.*



Counts: 1 | Confidence: From 0 To 0.8 | Remove Stop Words

Figure 6: *Find speech recognition errors in the well-known speech "I have a dream" by Martin Luther King Jr. 1963.*

words in a speech, whose size and color can be assigned according to different attributes. As shown in Fig. 4, the terms are shown when they appear multiple times and have a large entropy of audio lengths. A large word indicates that it is talked with very different lengths by the speaker(s), and the color a word indicates confidence. Here, the term "president" is clicked for drill-down study due to a large entropy. The scatter plot shows all instances, where the x-axis shows confidence and the y-axis audio length. Users can click any instance to highlight (in brown) and listen to the fragment.

Sentiment and Comparative Visualization: It is important to show the sentiments of speakers together with speaking characteristics. The sentiment scores between -5 to 5 can be mapped to fragment or word as shown in Fig. 5. Comparing different speeches to visually discover speakers' similarities and differences quickly and intuitively (see Fig. 5 for side-to-side views of two speeches). More details are discussed in the case studies (see Sec. 6).

6. Case Studies

Comparing two speeches: Two speeches are visualized for comparison in Fig. 5. They are the Civil Right Speech (2015) by President Obama (Fig. 5(A)) and the Speech at United Nations (2018) by President Trump (Fig. 5(B)). From the overview of fragments colored by sentiment scores, it can be seen that Trump has more positive talk than Obama since the fragment bars have more yellow-orange colors (see color legend Fig. 5(C)). But his talk also has very dark and negative fragments. The information can be discovered from the statistics in Fig. 5(F). In Fig. 5(D), Obama's speech has larger confidence values (green arrow) which show his talk is relatively clearer for the speech-to-text tool. Fig. 5(E) indicates that Trump tentatively uses longer audio terms (purple arrow) which may be slow and/or emphasized.

Find speech recognition errors: Users can define combined conditions over different attributes to investigate recognized fragments or terms. In the example of Fig. 6, the well-known speech "I have a dream" is transcribed by Google Speech to Text tool. It is a public speech delivered by American civil rights activist Martin Luther King Jr. on August 28, 1963. From the histogram views in Fig. 6(B), it can be seen that there exist some low confidence terms below 0.7. By selecting this range and find terms with a single appearance in the speech (Fig. 6(C)), several suspicious keywords are shown in Fig. 6(A). By checking it in a detail view, or listening to the audio error can be recognized, and later it can be corrected manually.

7. Conclusion and Discussion

Speech to text AI tools are prevalent while the results often need to be revised and investigated to discover errors and understand speech/speaker features. A set of visualizations are presented in this type of emerging data. It is useful for many post-processing tasks in a variety of applications. This paper does not present a complete design and user study for all potential visualization functions. However, we show a preliminary prototype in this short paper. In the future, it will be extended in several facets: (1) the visual design will be further improved with alternatives; (2) effective guided interaction is preferred; and (3) the audio signal processing attributes can be integrated. The system will also be evaluated by domain users with a formal user study.

Acknowledgement

The authors would like to thank the anonymous reviewers. This work was supported in part by the U.S. NSF under Grant 1739491.

References

[BMG*16] Burgoon J., Mayew W. J., Giboney J. S., Elkins A. C., Moffitt K., Dorn B., Byrd M., Spitzley L.: Which spoken language markers identify deception in high-stakes settings? evidence from earnings conference calls. *Journal of Language and Social Psychology 35*, 2 (2016), 123–157. 1

[BZK12] Bumbalek Z., Zelenka J., Kencl L.: Cloud-based assistive speech-transcription services. In *International Conference on Computers for Handicapped Persons* (2012), Springer, pp. 113–116. 1

[CFL13] Clark A., Fox C., Lappin S.: *The handbook of computational linguistics and natural language processing*. John Wiley & Sons, 2013. 1

[CRS05] Collins M., Roark B., Saraclar M.: Discriminative syntactic language modeling for speech recognition. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (2005), Association for Computational Linguistics, pp. 507–514. 1

[CWM19] Chou J.-K., Wang Y., Ma K.-L.: Privacy preserving visualization: A study on event sequence data. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 340–355.

[GMH13] Graves A., Mohamed A. R., Hinton G.: Speech recognition with deep recurrent neural networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2013), 6645–6649. arXiv:1303.5778, doi:10.1109/ICASSP.2013.6638947. 1

[Goo19] Google: Google Cloud Speech-to-Text. *https://cloud.google.com/speech-to-text/* (2019). URL: https://cloud.google.com/speech-to-text/. 1

[HDY*12] Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., Kingsbury B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine 29* (2012), 82 – 97. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=6296526{&}isnumber=6296521, arXiv:1207.0580. 1

[HKPC18] Hohman F. M., Kahng M., Pienta R., Chau D. H.: Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* (2018). doi:10.1109/TVCG.2018.2843369. 2

[HM15a] Hirschberg J., Manning C. D.: Advances in natural language processing. *Science 349*, 6245 (2015), 261–266. 1

[HM15b] Hossain M. S., Muhammad G.: Cloud-assisted speech and face recognition framework for health monitoring. *Mobile Networks and Applications 20*, 3 (2015), 391–399. 1

[JM14] Jurafsky D., Martin J.: *Speech and Language Processing*, vol. 3. 2014. arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004. 1

[JTL*20] Jiang N., Tian F., Li J., Yuan X., Zheng J.: Man: mutual attention neural networks model for aspect-level sentiment classification in siot. *IEEE Internet of Things Journal* (2020). 2

[KAKC17] Kahng M., Andrews P. Y., Kalro A., Chau D. H. P.: A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics 24*, 1 (2017), 88–97. 1

[KCK*18] Kwon B. C., Choi M.-J., Kim J. T., Choi E., Kim Y. B., Kwon S., Sun J., Choo J.: Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 299–309. 1

[KJFF15] Karpathy A., Johnson J., Fei-Fei L.: Visualizing and Understanding Recurrent Networks. *arXiv preprint arXiv:1506.02078* (2015). URL: http://arxiv.org/abs/1506.02078, arXiv:1506.02078. 2

[KK15] Kucher K., Kerren A.: Text visualization techniques: Taxonomy, visual survey, and community insights. *IEEE Pacific Visualization Symposium 2015-July* (2015), 117–121. doi:10.1109/PACIFICVIS.2015.7156366. 2

[KRS17] Kisler T., Reichel U., Schiel F.: Multilingual processing of speech via web services. *Computer Speech & Language 45* (2017), 326–347. 1

[LSL*17] Liu M., Shi J., Li Z., Li C., Zhu J., Liu S.: Towards Better Analysis of Deep Convolutional Neural Networks. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 91–100. doi:10.1109/TVCG.2016.2598831. 2

[LWC*18] Liu S., Wang X., Collins C., Dou W., Ouyang F., El-Assady M., Jiang L., Keim D. A.: Bridging text visualization and mining: A task-driven survey. *IEEE transactions on visualization and computer graphics 25*, 7 (2018), 2482–2504. 2

[Mic] Microsoft: Microsoft Azure Speech to Text. 1

[MLJ*14] Mou L., Li G., Jin Z., Zhang L., Wang T.: Tbcnn: A tree-based convolutional neural network for programming language processing. corr abs/1409.5718 (2014). *arXiv preprint arXiv:1409.5718* (2014). 1

[Muh15] Muhammad G.: Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system. *Cluster Computing 18*, 2 (2015), 795–802. 1

[Nie11] Nielsen F. A.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages* (2011), 93–98. 2

[NPM] NPM: AFINN-based sentiment analysis for Node.js. URL: https://www.npmjs.com/package/sentiment. 2

[RFFT17] Rauber P. E., Fadel S. G., Falcão A. X., Telea A. C.: Visualizing the Hidden Activity of Artificial Neural Networks. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 101–110. doi:10.1109/TVCG.2016.2598838. 2

[RLGW18] Ragni A., Li Q., Gales M., Wang Y.: Confidence Estimation and Deletion Prediction Using Bidirectional Recurrent Neural Networks. *IEEE Workshop on Spoken Language Technology* (2018). URL: http://arxiv.org/abs/1810.13025, arXiv:1810.13025. 2

[SGPR18] Strobelt H., Gehrmann S., Pfister H., Rush A. M.: LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 667–676. doi:10.1109/TVCG.2017.2744158. 2

[WNY19] Wang L., Niu J., Yu S.: Sentidiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* (2019). 2

[WPW*11] Wu Y., Provan T., Wei F., Liu S., Ma K.-L.: Semantic-preserving word clouds by seam carving. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 741–750. 2

[ZZ18] Zhang Q., Zhu S.-C.: Visual Interpretability for Deep Learning: a Survey. URL: http://arxiv.org/abs/1802.00614, arXiv:1802.00614. 2