JURNAL TEKNIK INFORMATIKA (JUNIKA)
ISSN xxxx - xxxx
Volume 1 Number 1, April 2024
https://ejurnal.uniyos.ac.id/index.php/yij-server

Jurnal Review : Subtitle Generation Pada Video Movie

Luhfita Tirta Swarga
Teknik Informatika, Universitas Yos Soedarso Surabaya; luhfitatirta@gmail.com
Nur Kurniasari
Teknik Informatika, Universitas Yos Soedarso Surabaya; nurkurniasari.nia@gmail.com
Adithya Marhaendra Kusuma
Teknik Informatika, Universitas Yos Soedarso Surabaya; adit.marhaendra@gmail.com

ABSTRACT

Video is one of the most popular multimedia that is often used by the public. So it is very important to make the information covered in the form of sound in the video that can be understood by the general public. So the most natural way lies in the use of subtitles on the screen below the video. Therefore it is important to research subtitle generation. From this research it has been found that in the context of subtitle generation between Hesham and Sridhar the best method is Hybrid from researchers [Sridhar, et al], compared to the S-Trailer method from researchers [Hesham, et al]. in addition, there are other studies that also prove that the Latent Semantic Analysis (LSA) method can also be used for subtitle generation.
Keywords*: Subtitle Generation; Movie; Automatic Speech Recognition; Natural Language Processing*

ABSTRAK

Video merupakan salah satu multimedia yang paling populer yang sering digunakan oleh masyarakat . jadi sangat penting untuk membuat infromasi yang dicakup dalam bentuk suara dalam video yang dapat dipahami oleh masyarakat umum. Sehingga cara yang paling alami terletak pada penggunaan teks terjemahan dilayar bawah video. Oleh karena itu penting untuk meneliti subtitle generation. Dari riset ini telah ditemukan bahwa dalam konteks subtitle generation antara Hesham dan Sridhar metode yang paling bagus adalah Hybrid dari peneliti [Sridhar, et al], dibandingkan dengan metode S-Trailer dari peneliti [Hesham, et al]. selain itu, ada penelitian lain yang juga membuktikan bahwa metode Latent Semantic Analysis (LSA) juga bisa digunakan untuk subtitle generation.
Kata kunci: Subtitle Generation; Movie; Automatic Speech Recognition; Natural Language Processing

PENDAHULUAN

Latar Belakang

Dengan tingkat pertumbuhan meluasnya ketersediaan internet berkecepatan tinggi menyebabkan video menjadi media informasi yang akrab disitus website repositori video seperti google, dailymotion, vimeo, youtube, dan lain-lain[1] yang luar biasa dalam video yang dibuat pengguna, dan menjadi semakin penting untuk dapat menavigasi mereka secara efesien[2].

Video merupakan salah satu multimedia yang paling populer yang sering digunakan oleh masyarakat. Jadi sangat penting untuk membuat informasi yang dicakup dalam bentuk suara dalam video yang dapat dipahami oleh masyarakat umum. Cara yang paling alami terletak pada penggunaan teks terjemahan[3]. Subtitle merupakan terjemahan teks dari dialog dalam video yang ditampilan secara real time selama pemutaran video dibagian bawah layar[4]. Namun, pembuatan subtitle manual merupakan kegiatan yang sangat membosankan dan membutuhkan partisipasi aktif pengguna. Proses mengenali secara otomatis kata-kata yang diucapkan pembicara berdasarkan informasi dalam sinyal ucapan disebut pengenalan suara[5].

Pengenalan suara merupakan terjemahan kata-kata yang diucapkan ke dalam teks. Itu juga dikenal sebagai Automatic Speech Recognition (ASR)[6], Computer Speech Recognition atau Speech To Text (STT). Natural Language Processing (NLP) merupakan suatu bidang ilmu computer yang berhubungan dengan interaksi antara computer dan Bahasa manusia, ini berkaitan dengan membuat computer untuk memahami dan menafsirkan bahasa manusia[7]. Dalam pengenalan suara otomatis, computer menangkap kata-kata yang diucapkan oleh manusia dengan bantuan mikrofon, kata-kata ini

kemudian dikenal oleh pengenal ucapan otomatis dan pada akhirnya system menampilkan kata-kata yang dikenal di layar bawah video. Sistem pengenalan suara otomatis real time untuk video menghadapu tantangan yang besar untuk meningkatkan akurasi dan kecepatan pengenalan, dan kinerja sistem pengenalan suara menurun karena kebisingan dan ketepatan saat berbicara. Ini juga akan dipengaruhi oleh data video yang bervariasi karena ketergantungan pada jenis kelamin pembicara serta kondisi lingkungan dan gaya bicara[5].

Berdasarkan survey jurnal oleh[2] terkait dengan S-Trailer : Automatic Subtitle Generation ditemukan beberapa pendekatan terbaik, pendekatan yang dilakukan Hesham menggunakan Teknik Natural Language Processing (NLP), Machine Learning, dan Deep Learning untuk menangkap pendapat dan perilaku pengguna, agar mampu memberikan rekomendasi adegan video yang relevan sesuai dengan prefensi pengguna. Survey jurnal bertujuan mencari metode pendekatan dan algoritma subtitle generation berbahasa Indonesia di video movie trailer.

METODE

Jurnal ini tersusun secara metodologi melalui proses yang terdiri dari lima tahapan, yaitu: Penentuan Kriteria Jurnal, Penentuan Sumber, Pemilihan Literatur, Pengumpulan Data, dan Pemilihan Data.

A. Tahap 1 : Penentuan Kriteria Jurnal
Ditentukan berdasar Inclusion Criteria (IC) :
1) IC1 : Jurnal harus berupa penelitian asli
2) IC2 : Jurnal terbit antara 2015 sampai 2022
3) IC3 : Jurnal yang diteliti tentang subtitle generation

Jurnal yang diambil diberi batasan keaslian tahun penerbit dan topik yang memberikan informasi metode dan hasil temuan yang cukup actual mengenai topik yang relevan.

B. Tahap 2 : Penentuan Sumber
Google Scholar dan IEEE digunakan sebagai sumber pengambilan literatur dikarenakan hal ini terkait juga pada kualitas jurnal yang menjamin keasliannya dan juga untuk memberikan kemudahan dalam pencarian jurnal.

C. Tahap 3 : Pemilihan Literatur
1) Kata kunci yang digunakan adalah "*Subtitle Generation*" dan "*Automatic Speech Recognition*".
2) Abstrak dari hasil pencarian jurnal dibaca sebagai penentu pemilihan jurnal, apakah diterima atau ditolak.

D. Tahap 4 : Pengumpulan Data
Pencarian dan pemilihan jurnal dilakukan secara manual. Dari kata kunci "*Subtitle Generation*" dan "*Automatic Speech Recognition*", didapat 8 jurnal dan yang terpilih 3 jurnal tentang Subtitle Generation berbahasa asing.

Tabel 1. Pengumpulan Data

Sumber	Kandidat	Terpilih
Google Scholar	6	1
IEEE	2	2

E. Tahap 5 : Pemilihan Data
Jurnal yang digunakan adalah jurnal dengan penelitian yang berisi metode pendekatan subtitle generation di movie trailers.

JURNAL TEKNIK INFORMATIKA (JUNIKA)
ISSN xxxx - xxxx
Volume 1 Number 1, April 2024
https://ejurnal.uniyos.ac.id/index.php/yij-server

HASIL DAN DISKUSI

A. Dataset

Jurnal yang terpilih pada umumnya mengambil sekumpulan video di internet dengan memberikan label secara manual sebagai dataset.

Tabel 2. Dataset

Penulis	Tahun	Dataset
Sridhar et al	2014	5000 film dari IMDB di kaggle
Hesham et al	2018	Tidak ada dataset
Aswin et al	2019	40 Video dengan durasi yang bervariasi

B. Kontribusi Pendekatan dan Hasil

Berikut ini merupakan kontribusi dan metode-metode pada jurnal yang terpilih.

Tabel 3. Kontribusi dan Pendekatan

Penulis	Tahun	Kontribusi	Pendekatan
Sridhar et al	2014	Discourse segment detection (DSD) dengan hybrid model dengan akustik dan linguistic	Discourse segment detection (DSD), Acoustic pause, POS statistics
Hesham et al	2018	S-Trailer framework yaitu merupakan model natural language processing yang terintegrasi dengan teknik machine learning	Bag of words, NLP (natural language processing), machine learning, deep learning
Aswin et al	2019	Automatic subtitle generation dan teknik semantic video summarization	Metode In-tersection dan metode weight based learning

Pada penelitian Hesham[2], Bag of words, Natural language processing (NLP), Machine learning, dan Deep learning digunakan untuk pembuktian validitas framework yang diusulkan melalui cara Bag of words yang dihasilkan dari hasil uji terhadap sejumlah film di kaggle. Ada sekitar 5000 film dari genre yang berbeda dan disetiap dataset dalam film disediakan dengan genre yang terkait dengan IMDB.

Awal dari framework yang diusulkan adalah serangkaian daftar yang berisi sekumpulan kata dan frasa yang berbobot dimana disetiap aliran film melalui pemanfaatan Bag of words yang diusulkan framework yang dapat mengklasifikasi subtitle genre film. Pada tabel 4 berikut menyajikan akurasi klasifikasi berbagai genre film yang dihasilkan dari framework yang diusulkan.

JURNAL TEKNIK INFORMATIKA (JUNIKA)
ISSN xxxx - xxxx
Volume 1 Number 1, April 2024
https://ejurnal.uniyos.ac.id/index.php/yij-server

Tabel 4. Evaluasi Akurasi Klasifikasi S-Trailer

Genre	Classification Accuracy
Action	89%
Comedy	75%
Drama	72%
Romance	60%
Fantasy	55%

Dalam percobaan peneliti mengurutkan genre yang diperhitungkan pertimbangannya ketika peneliti ingin mengevaluasi klasifikasi ketepatannya. Misalnya, jika genre filmnya adalah Action, Crime, dan Romance. Classifier harus mengembalikan urutan yang sama pada penampilan genre film yang dianggap sebagai kesalahan klasifikasi.

Hasilnya yaitu dapat dianggap sebagai seed corpus untuk klasifikasi film, maka tidak ada standart golden corpus yang digunakan untuk mengklasifikasi film.

Evaluasi trailer yang dihasilkan akan didasarkan pada Precision, Recall, dan F-Measure Metrics :

Dimana PRECISION adalah rasio jumlah adegan yang relevan yang diambil dari jumlah total adegan trailer asli[8], [9], yang dirumuskan sebagai berikut :

$$Precision = \frac{Number\ of\ relevant\ scenes\ retrieved}{Total\ number\ of\ original\ trailers\ scenes} \quad (1)$$

RECALL adalah rasio jumlah adegan yang benar atau relevan yang diambil dari jumlah total adegan yang relevan, lalu diindeks dalam database trailer yang dihasilkan[9], yang dirumuskan sebagai berikut :

$$Recall = \frac{Number\ of\ relevant\ scenes\ retrieved}{Total\ number\ of\ generated\ trailer\ scenes} \quad (2)$$

F-MEASURE adalah rata-rata harmonic dari precision dan recall[9], yang dirumuskan sebagai berikut :

$$F - measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (3)$$

Untuk tujuan evaluasi, 50 film digunakan untuk mengevaluasi kinerja model yang diusulkan. Tabel V menunjukkan rata-rata kinerja S-Trailer dalam pengambilan adegan yang serupa ditampilkan di trailer asli.

Tabel 5. Evaluasi Kinerja S-Trailer

Genre	Average Precision	Average Recall	Average F-measure
Action	0.38	0.60	0.47
Comedy	0.35	0.42	0.38
Drama	0.44	0.46	0.45
Romance	0.31	0.34	0.32
Fantasy	0.26	0.30	0.32

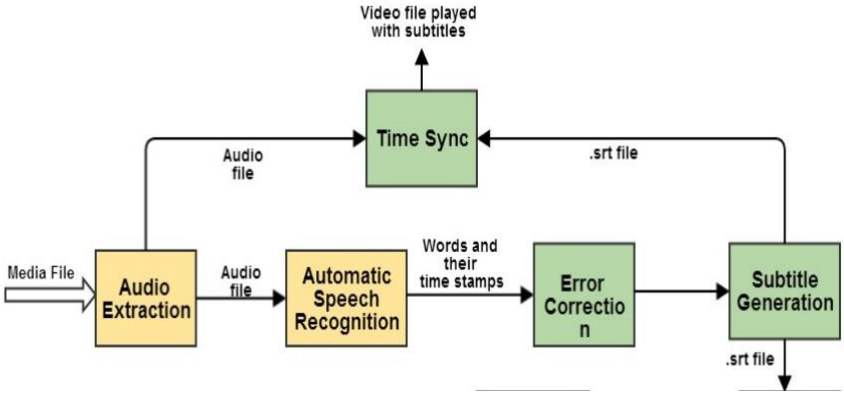
Hasil eksperimen menunjukkan hasil yang menjanjikan oleh S-Trailer. Kelemahan utama dari metode yang diusulkan yaitu bahwa tidak bisa mengambil adegan dimana ada dan tidak adanya pidato

JURNAL TEKNIK INFORMATIKA (JUNIKA)
ISSN xxxx - xxxx
Volume 1 Number 1, April 2024
https://ejurnal.uniyos.ac.id/index.php/yij-server

didalamnya. Peneliti mencatat bahwa produser dari trailer asli cenderung menggunakan adegan diam seperti itu untuk tujuan ketertarikan. Namun, kelemahan tersebut akan dipangkas saat dimasa depan.

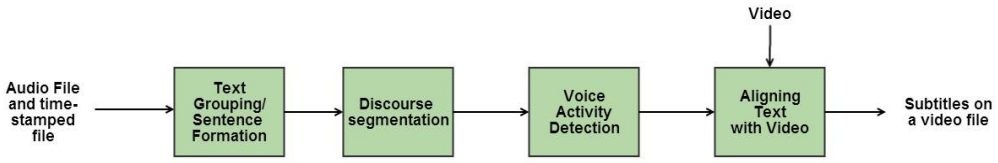
Kesimpulan dari penelitian ini yaitu untuk menunjukkan bagaimana modelnya dapat berhasil digunakan diseluruh fase bidang pemasaran film untuk mengekstrak trailer yang menarik untuk penonton.

Berikut ini adalah penjelasan algoritma pada penelitian Sridhar[7] pada gambar 1 arsitektur blok diagramkeseluruhan menjelaskan bahwa audio diekstraksi dari media file audio menggunakan opsi "Konversi" di VLC Media Player untuk mendapatkan tipe file .wav. file audio yang sudah diproses lalu diteruskan ke model Automatic Speech Recognition (ASR). Modul ASR mengimplementasikan CMUSphinx API. Kegunaan Sphinx API yaitu untuk mengambil pra-proses file audio dan memberikan transkrip yang sesuai dengan waktu file audio yang akan digunakan untuk subtitle generation selanjutnya. Mesin pengenalan suara CMU Sphinx memerlukan 2 jenis file untuk mengenali suara yaitu model akusti8k dan model bahasa. Model akustik dibuat untuk merekam data lisan dan transkrip mereka, lalu mengkomplikasikannya menjadi representasi statistic dari unit suara yang membentuk setiap kata. Model bahasa yaitu menangkap property bahasa dan memprediksi kata berikutnya dalam urutan ucapan.



Gambar 1. Block Diagram

Discourse Segment Detection (DSD) adalah bagian penting dari subtitle generation yang melibatkan pendekatan hybrid yang di dapat dari pendekatan linguistic dan akustik. Modul ini menemukan akhir dari setiap baris subtitle yang menggunakan akustik dan dalam kasus kalimat yang panjang pendekatan linguistic lebih banyak digemari.



Gambar 2. Discourse Segment Detection

ASR menghasilkan string sebesar kata-kata dan peneliti mengaitkan waktu dengan mereka. String tidak tersegmentasi dengan benar ke dalam discourse segment dank arena itu sulit dimengerti dan juga ditampilkan pada waktu dilayar internet. Teknik DSD dapat digunakan dimana batas segment didalam sistem, peneliti menggabungkan pendekatan hybrid yang menggunakan fitur akustik dan fitur linguistic dimanapun diperlukan. Fitur akustik diberikan prioritas lebih tinggi karena ada pembatasan pada panjang setiap baris subtitle. Jadi kapanpun banyak jeda akustik terjadi itu diambil sebagai

JURNAL TEKNIK INFORMATIKA (JUNIKA)
ISSN xxxx - xxxx
Volume 1 Number 1, April 2024
https://ejurnal.uniyos.ac.id/index.php/yij-server

istirahat dan garis subtitle dihentikan pada titik waktu tersebut untuk memenuhi format standart .srt. fitur linguistic berdasarkan pada bagian dari data lisan disetiap batas dan dengan demikian dapat mendeteksi batas kalimat dengan probabilitas maksimum menjadi akhir discourse segment.

Semua data lisan dalam discourse dibagi menjadi 5 kategori menurut Grosz dan Hirschberg, sebagai berikut :

1. Segment initial sister (SIS)
2. Segment initial embedded (SIE)
3. Segment medial (SM)
4. Segment medial pop (SMP)
5. Segment final (SF)

Dalam pendekatan peneliti, 2 kategori pertama yaitu SIS dan SIE digabungkan menjadi satu kategori segment beginning utterances (SBEG). SMP dan SF digabungkan sebagai ucapan terakhir (SFIN). Dalam pendekatan peneliti, peneliti mendeteksi batas dengan probabilitas maksimum SFIN yang diikuti oleh SBEG.

Berikut ini beberapa hasil dari penelitian ini sebagai berikut :

1. Pada tabel 6 menunjukkan jumlah dan kategori berbagai discourse boundaries yang ditemui setelah pengenalan suara.

Tabel 6. Hits Vs Total Utterances

Category	Number of Hits	Total in Sample
SBEG	15	43
SFIN	5	147
Totals	**22**	**197**

2. Pada tabel 7 menunjukkan nilai dari berbagai parameter untuk discourse segmentation menggunakan pendekatan peneliti dan pendekatan batas yang ada.

Tabel 7. Comparison between Existing Boundary and Proposed Hybrid Approach

	Recall	Precision	fallout	Error
Hybrid Approach	0.35	0.68	0.05	0.18
Boundary Approach	0.25	0.82	0.03	0.30

Pada penelitian Aswin[1], dataset berupa 40 video dengan waktu yang bervariasi. Video-video ini diberikan kepada 4 algoritma ini serta juga algoritma gabungan dan jumlah baris yang diperoleh dalam video yang dirangkum. Effisiensi masing-masing algoritma ditentukan berdasarkan output dari video gabungan yang merupakan persimpangandari semua algoritma. Jadi, efisiensi dapat didefinisikan untuk suatu algoritma sbg perbandingan jumlah subtitle dlm video gabungan dengan output dari algoritma tertentu.

$$Efficiency = \frac{Ncombined}{Nalgorithm} \quad (4)$$

a. Efficiency of video summarization (efisiensi peringkasan video)

Dengan peringkasan dalam suatu algoritma ini, berarti berapa banyak bagian dari algoritma yang dirangkum hadir dalam video yang dihasilkan oleh teknik ensemble. Seperti rumus yang disebutkan diatas bahwa efisiensi setiap algoritma dapat dihitung dengan menggunakan rumus diatas dan pada penerapan hasil metode persimpangan berikut ini diperoleh seperti yang ditunjukkan pada tabel berikut :

JURNAL TEKNIK INFORMATIKA (JUNIKA)
ISSN xxxx - xxxx
Volume 1 Number 1, April 2024
https://ejurnal.uniyos.ac.id/index.php/yij-server

Tabel 8. Efficiency of Intersection Method

Lex Rank	LSA	Luhn	Text
37.1	40.6	38.0	37.7

Pada penerapan dataset teknik weighted ensemble, dengan bobot awal diatur ke 1 untuk semua 4 algoritma demikian berikut hasil yang diperoleh seperti yang ditunjukkan pada tabel 9 dan bobot yang dperbarui ditunjukkan pada tabel 10.

Tabel 9. Efficiency of Weighted Ensembel Method

Lex Rank	LSA	Luhn	Text
18.5	61.0	38.0	37.7

Tabel 10. Weighted Ensembel Method

	LSA	Luhn	Text	Lex Rank
Initial	1	1	1	1
Finial	1.975	1	1	0.025

Dari tabel 8 dan 9 telah diamati bahwa LSA berkinerja lebih baik dan bentuk Lex Rank yang paling sedikit. karena untuk semua video LSA memiliki kinerja terbaik dan Lex memiliki bobot paling rendah sedangkan bobot Luhn dan Text Rank tetap tidak berubah (Tabel 10). Ada perbedaan besar dalam nilai efisiensi metode berbasis berat dan persimpangan karena dalam teknik ensemble berbasis berat pada setiap iterasi, bobot algoritma meningkat menjadi lebih baik dan algoritma terburuk mengalami penurunan. Dari kedua metode ini telah menjelaskan bahwa LSA memiliki kontribusi besar terhadap video yang dirangkum dan Lex memilik kontribusi paling sedikit.

b. Complexity (Kompleksitas)

Kompleksitas dibagi menjadi 2 bagian sebagai berikut :

- Kompleksitas Waktu

1. Algoritma Summarization Tunggal :

$$\mathbf{0}(\boldsymbol{nk})$$

Dimana *n* adalah jumlah iterasi yang diperoleh *n* sampai panjang summarization summarization dan *k* adalah jumlah kalimat dalam subtitle yang diringkas.

2. Algoritma Summarization Gabungan :

$$\mathbf{0}\sum_{t=1}^{\propto}\boldsymbol{n_i k_i} + \boldsymbol{min}\,(\boldsymbol{k_i})$$

Dimana α adalah jumlah metode yang akan digabungkan, *n* adalah jumlah iterasi yang diperoleh sampai panjang summarization dan *k* adalah jumlah kalimat dalam subtitle yang diringkas.

- Kompleksitas Ruang

1. Algoritma Summarization Tunggal :

$$\mathbf{0}\,(\boldsymbol{r} + \boldsymbol{L} * \boldsymbol{r})$$

Dimana *r* adalah jumlah total wilayah dalam array subtitle dan *L* adalah panjang rata-rata kalimat dalam subtitle yang diringkas.

2. Algoritma Summarization Gabungan :

$$0 \sum_{k=1}^{\infty} r + L * r$$

Dimana *r* adalah jumlah total wilayah dalam array subtitle dan *L* adalah panjang rata-rata kalimat dalam subtitle yang diringkas.

KESIMPULAN

Dalam konteks subtitle generation, antara Hesham dan Sridhar metode yang paling bagus adalah metode hybrid dari peneliti Sridhar[7] karena nilai precisionnya 0.68 dibandingkan dengan metode S-Trailer dari penelitian Hesham[2] yang nilai precisionnya hanya 0.34.

Selain itu, ada penelitian lain yang juga membuktikan bahwa metode Lantet Semantic Analysis (LSA) juga bisa digunakan untuk subtitle generation[1].

DAFTAR PUSTAKA

[1] A. VB *et al.*, "NLP Driven Ensemble Based Automatic Subtitle Generation and Semantic Video Summarization Technique," *arXiv Prepr. arXiv1904.09740.*, 2019.

[2] M. Hesham, B. Hani, N. Fouad, and E. Amer, "Smart Trailer : Automatic generation of movie trailer using only subtitles," *2018 First Int. Work. Deep Represent. Learn.*, pp. 26–30, 2018.

[3] K. Mishra, P. Bhagat, and A. Kazi, "Automatic Subtitle Generation for Sound in Videos," in *International Journal of Engineering and Technology (IRJET)*, 2016, vol. 3, no. 2, pp. 915–918.

[4] A. Jakhotiya, K. Kulkarni, C. Inamdar, B. Mahajan, and A. Londhe, "Automatic Subtitle Generation for English Language Videos," in *International Journal of Computer Science and Engineering*, 2015, vol. 2, no. 10, pp. 5–7, doi: 10.14445/23488387/ijcse-v2i10p102.

[5] R. S. Chavan and G. S. Sable, "An Implementation of Text Dependent Speaker Independent Isolated Word Speech," *Int. J. Eng. Sci. Res. Technol.*, vol. 2, no. 9, 2013.

[6] L. Tirta, J. Santoso, and E. Setyati, "Pengenalan Lirik Lagu Otomatis Pada Video Lagu Indonesia Menggunakan Hidden Markov Model Yang Dilengkapi Music Removal," *J. Inf. Syst. Hosp. Technol.*, vol. 4, no. 2, pp. 86–94, 2022, doi: 10.37823/insight.v4i2.225.

[7] R. Sridhar, S. Aravind, H. Muneerulhudhakalvathi, and M. Sibi Senthur, "A hybrid approach for Discourse Segment Detection in the automatic subtitle generation of computer science lecture videos," *Proc. 2014 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2014*, pp. 284–287, 2014, doi: 10.1109/ICACCI.2014.6968422.

[8] N. Nazari and M. Mahdavi, "A survey on Automatic Text Summarization," *J. AI Data Min.*, vol. 0, no. 0, pp. 121–135, 2018, doi: 10.22044/jadm.2018.6139.1726.

[9] S. Liu, "CS585 Project Report Long Text Summarization using Neural Networks and Rule-Based Approach," 2017.