International Journal of Modern Agriculture, Volume 9, No.3, 2020
ISSN: 2305-7246

Impact of Positive Journalism: A Machine Learning Perspective

Nikhil Matta[1], Sachin Pondichurri[2], Sagarika Jain[3], Ujjwal Jain[4], Vishnu Vellure[5], Preeti Mulay[6], Rahul Joshi[7]
[1,2,3,4,5,6,7] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[6]Email: preeti.mulay@sitpune.edu.in[6]

Abstract

With the exponential rise and change in technology over recent times, the way users consume data has changed along with the changes in technology. The data we consume on a regular basis has a direct correlation to how we feel and react. Dr APJ Abdul Kalam, the 11th president of India, once stated; "Why is the Indian media so negative? Why are we in India so embarrassed to recognize our own strengths, our achievements? We are such a great nation. We have so many amazing success stories but we refuse to acknowledge them. There are millions of such achievements but our media is only obsessed with the bad news and failures and disasters." With the direction in which the media is headed all over the world, we decided to try and comprehend its effects on readers.

In this paper, we attempt to understand how the news that we regularly go through affects us and what role machine learning could play to help the situation. The work done in this paper can be used to help further the research in the field of natural language processing and its use in news. Since the advent of internet technologies, databases and libraries of e-News Paper is getting built over the years. Few university libraries and online libraries / archives become very useful data source for researchers from media domain. These huge data sources are useful for researchers from IT domain too for applying various classification techniques to know further details and insights which are printed as news, published worldwide, read by many and used too.

Key words: positive news, positive journalism, natural language processing, newspaper, machine learning, NLP

Introduction

Over the course of the past few years, it has been seen that though there may be a minor decline in online news(due to concerns regarding fake news), there has at the same time been an increase in the process of people sharing news through social media. Along with these, there has also been a significant increase in the percentages of people paying specifically for getting their news online[1]. This basically gives us an understanding that online news has been slowly coming up to be a primary source of news for a lot of people all over the world. The accessibility that the internet provides could also reflect on the fact that it leads to people being more exposed to everything that is happening around them constantly. This may have some impact on how a person thinks and feels. In a paper written by Stuart N. Soroka[2], it is seen that users tend to be impacted more by negative news in terms of how they feel about it. In this paper, we attempt to see what impact positive news would have on the average internet user when exposed to it on a regular basis and how machine learning would fit into this situation.

For this purpose, we have created a system on the basis of machine learning algorithms that effectively classifies news into negative news and 'other' news. This helps us eliminate news that is blatantly negative while still being able to provide information regarding negative subjects from a comparatively neutral point of view while focusing more on positive news. For this purpose, we have created a dataset of 2000 news articles obtained from various sources during

the month of January 2020. We intend to make this dataset public as it may be beneficial for any future projects done by any other researchers or students.

Another point to observe is that due to the increase in social media usage across the spectrum of age groups, people now prefer to get information in quick hits rather than a long and drawn-out process of finding information, understanding it and processing it. This is quite contrasting to the way in which traditional news outlets provide news. In this context, we use the term 'traditional news outlets' to represent news media houses that have initially started out as physical newspapers and later over time, moved online along with traditional newspapers being in circulation. A lot of these websites still provide news in the form of a long article spanning over 3-4 paragraphs. Based on some recent data, the average internet user would pay attention while reading an article on the internet for around 15 seconds[3]. This brings us to the second part of this paper, which is, summarization.

Over recent years, there has been an increase in services online that would summarise articles and provide them to users regularly so they get the information they need in short and quick bursts. This is still(for the most part) a system that is run manually. As in, there are editors working round the clock manually making summaries of news articles found on various traditional sources. This brings us to one of the main aspects of this paper, that is to create a machine learning model that can handle this job and thereby completely automating the process of summarization of these news articles. In the following sections, we see how all of this would work, in detail.

System Architecture and Functioning

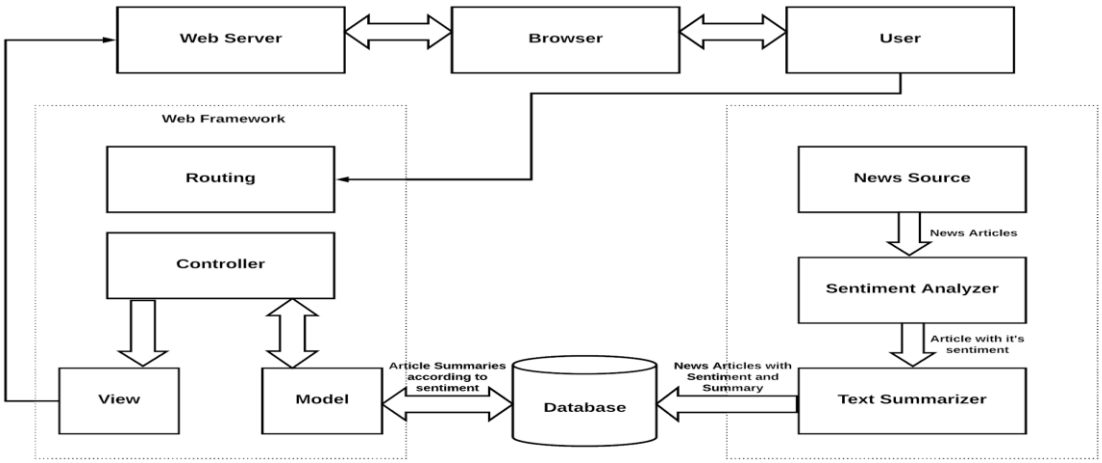

Fig 1. General system architecture

General Functioning

From *Fig. 1*, we get a clear idea of how the entire system functions. The system is designed to regularly collect articles from various fixed news sources. The data that is collected is all stored in a database. From this database, the collected articles are then analysed according to sentiment and the result of this analysis is now appended to the articles in the database. After this, the articles are subjected to summarization and stored in the database along with the respective article. From here, the data is displayed on the basis of positivity in sentiment on a website that we created for users to access.

For the most part, the entire system can be broken down into 3 broad phases as follows:

1. Collection of required data
2. Sentiment analysis model
3. Text summarization model

Along with these tasks, we have the tasks of creating and maintaining a NoSQL database and a user-friendly website.

Collection of Required Data

This is the first step to the functioning of this project. In order to create or rather tune a model to work we need data to train it on. For this purpose, we have collected a set of 2000 articles from various online news media outlets. Training the models on news articles specifically rather than general English language corpora would enable the model to gain a better level of accuracy in terms of grasping the context and the nature of the text.

This phase also involves the process of labelling the collected training data as ‘negative’ articles and ‘other’ articles. The idea behind this is to enable the model to filter out articles that may be completely negative and deliver articles that are either completely positive or have stronger positive connotations as compared to negative.

To do this we have manually labelled each of the 2000 articles as ‘negative’ and ‘others’. The basis on which this has been done is very close to what has been described by Aashish Agrawal et al.

Negative: If the subject of the article is a war, accidents, disaster, epidemic disease or killing, criminal activities, the death of a famous or important person, some sort of discrimination, bullying or stereotypes, some negative influence or event regarding economics.[4]

Other: If the subject of the article is anything that can not be termed as ‘negative’ then it can fall into this category

The purpose of this is to completely eliminate negative news and have a broader set of news articles that are either positive or close to positive. The reason this is necessary is to have a

constant stream of articles that are ready for the user to view. Also, based on the news articles of 2018, we see that 80.9% of news is negative and 19.1% of news is either positive or not negative. This is also the reason as to why we consider neutral articles as positive as well.

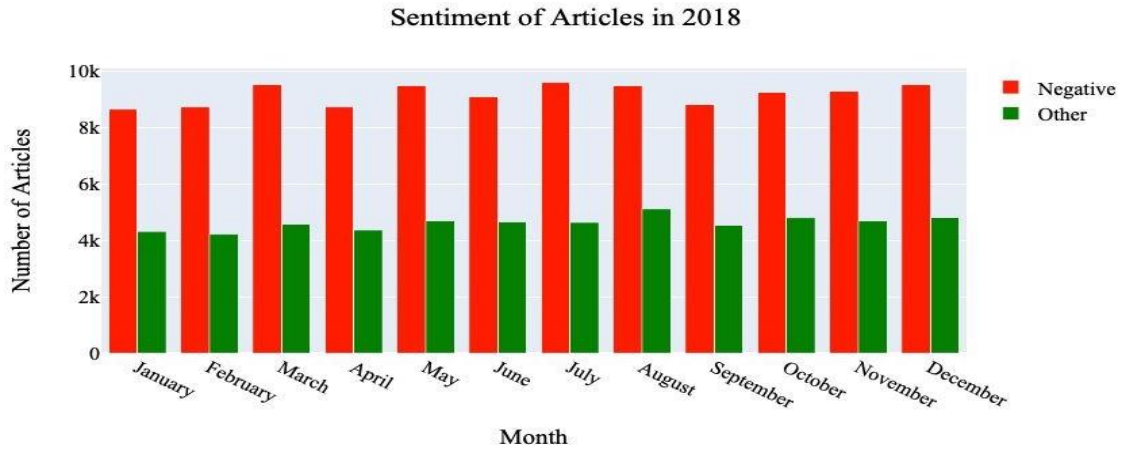

Fig 2. Article Count per month in terms of 'negative' and 'others' in 2018.

Month	**Sentiment**	
	Negative	**Other**
January	8642	4311
February	8722	4209
March	9529	4582
April	8722	4360
May	9489	4699
June	9090	4655
July	9588	4632
August	9492	5109
September	8825	4533
October	9255	4796
November	9283	4685
December	9526	4803

Fig 3. Monthly Article count in detail in the year 2018

Sentiment Analysis Model

After the collection of required data, we used the dataset to train our sentiment analysis

model. We tried various approaches for this purpose starting with NLTK's Vader Sentiment Analyzer. But since it just classifies text according to its sentiment using a lexicon of positive and negative words, we did not get good accuracy. After this, we tried out TextBlob, a python library to process textual data. But again the result for sentiment analysis of news articles was unsatisfactory. Ultimately we decided to tackle this problem by using transfer learning. Transfer Learning is using a pre-trained model on a new problem by just fine-tuning it for domain adaptation. The pre-trained model is generally trained on a huge amount of data and the internal layers are saved. These pre-trained models can then be fine-tuned to adapt to a specific domain.

We initially used Google's BERT as our pre-trained model. Adding a layer of classifier on BERT for sentiment analysis using our labelled news polarity dataset gave us much better results than any of our previous approaches. Later to further improve the accuracy of our model we tried out RoBERTa, an optimized BERT Pre-training Approach. RoBERTa was trained by Facebook AI on a much larger data corpus and for much longer than Google's BERT[5]. RoBERTa is trained on an additional 76GB of CC-News data and thus this was our primary reason to choose RoBERTa as the encoder for our model as it has been optimised on a large amount of news data. We added a classifier layer on top of RoBERTa to adapt it to our task of sentiment analysis of news articles by using our labelled news polarity dataset. After testing the accuracy, we found that RoBERTa understands news data better than BERT and thus gave us a better accuracy to classify the articles according to their sentiment.

To implement transfer learning using RoBERTa, we used Transformers, a State-of-the-art Natural Language Processing library developed by HuggingFace Inc[6]. The default arguments were used to train the model for sentiment analysis.

Comparative Analysis of Accuracy	
RoBERTa	90%
BERT	86%
Stanford CoreNLP	78%
TextBlob	73%
NLTK Vader	71%

Fig 4: Sentiment Model Accuracy Comparison

From *Fig. 4* we can see that the RoBERTa model is by far the most accurate in terms of predicting sentiments. We also noticed that some inaccuracy creeps in due to the model not knowing a few words, like 'coronavirus'.

Text Summarization Model

There are two basic approaches to text summarization:-

(i) Extractive Text Summarization

(ii) Abstractive Text Summarization

The extractive method creates a summary using words and phrases from the original text. Whereas the abstractive method forms more human-like summaries by paraphrasing the original text. We tried out various extractive and abstractive methods for our use case and found that abstractive summarization is better suited for news articles. Thus we used BART (Bidirectional & Auto-Regressive Transformers), a denoising autoencoder for pre-training sequence to sequence models developed by Facebook AI[7], to implement abstractive summarization. BART has been fine-tuned on the CNN/DailyMail dataset to produce summaries of news articles.

Experiment

To understand how well the system's outputs are, compared to the summaries of existing products carrying out the task of summarization and to validate the system's accuracy, we carried out an experiment. The idea of this experiment is to establish that the system that we have assembled here produces summaries that are close to those generated by humans.

For this experiment, we calculated ROGUE scores of the article and its human-generated summary and then the ROGUE scores of the article and the summary of the same article generated by the system that has been created by us. This lets us understand how close both the summaries are while keeping the same complete article as a standard common point between the two summaries.

We first calculated the ROGUE-N1 scores of the computer(BART) generated summaries with reference to the actual article which it is a summary of. Similarly, ROGUE-N1 results were also calculated for the summaries that were generated manually with reference to the actual article which is a summary of. The outcomes of this experiment are given below.

Results

Summarizer	**Recall**	**Precision**	**F-Measure**
BART	0.166519	0.986932	0.27164
Manually written summaries	0.152956	0.811858	0.247171

Fig 5. Outputs of the validation experiment

In *Fig. 5,* Recall, in the context of ROGUE, refers to how much of a reference summary a summary created by the computer can recover. Precision refers to what part of the summary created by the system is actually relevant or needed in a comparative viewpoint. Due to these scores being calculated with the full article as a common reference point, the Recall scores are drastically reduced while Precision scores are improved. BART achieving higher scores in both domains shows that the experiment here works and that it is highly capable of delivering summaries while maintaining the original content and context of the source article.

The F measure is the harmonic mean of Recall and Precision values and is used for comparison.

Survey

To get a better understanding of how this would perform as a product as compared to traditional news sources, we have managed to conduct a survey amongst a group of 100 people. For this purpose, we created our own digital newspaper in the format of a traditional paper. We manually picked and arranged the most relevant article summaries from all the summaries generated by the system. This was done every day for a week. The generated newspapers were sent along with a digital copy of a traditional paper to a selected group of individuals from a wide-ranged age group.

The purpose of this was to show the contrast in the type of content and how it was presented in traditional news as compared to the way we could potentially do. This was done for a week which was then followed by a survey. The idea of the survey was to get an idea as to how users respond to getting more news that is not negative as compared to the sensationalism present in traditional news sources. This brought us a variety of responses that can be seen in the next section.

Results of the survey

Based on the survey, we summarised the results into a more comprehensive form and presented it in this section. The following set of graphs show the responses to the questions that we have asked a group of 100 people in an attempt to understand how the general public responds to the

type of news that is presented to them. We also attempt to see how the type of information is pervasive in their personal lives as well.

1. *Fig. 6* shows that 52% of the participants said they felt negatively impacted by the amount of negative news in media tends to affect their mindset

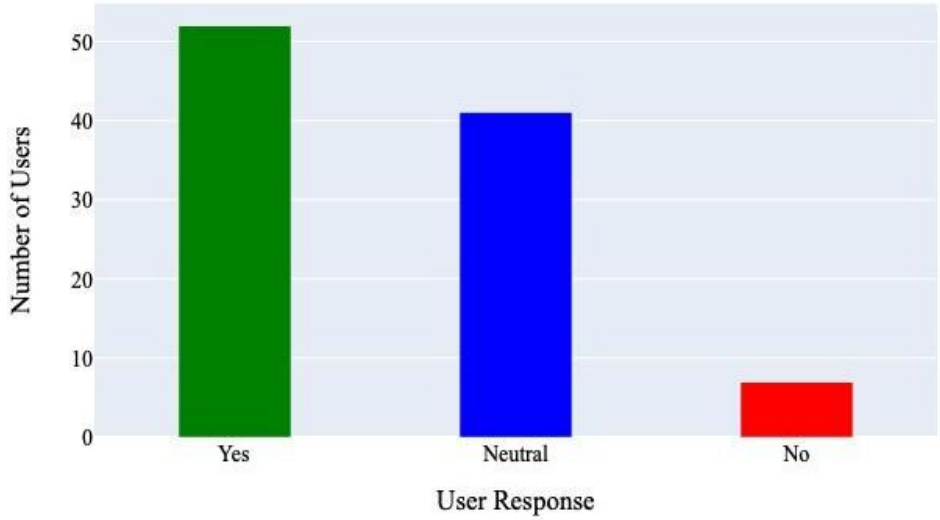

Fig 6. Impact of constant exposure to negative news

2. *Fig. 7* shows that approximately 75% of the participants prefer reading negative news when preceded by non-negative news

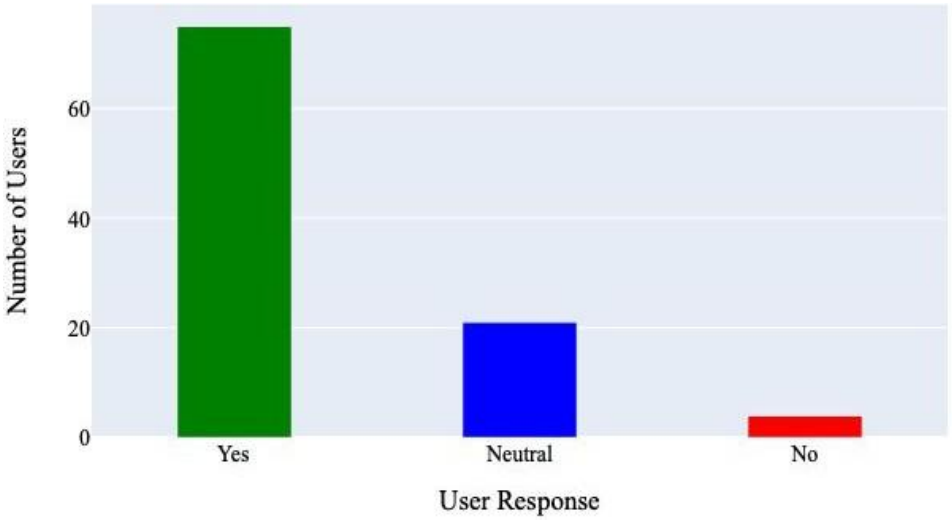

Fig 7. Users preferring positive news to appear first

3. *Fig. 8* shows that 66% of the participants tend to generally talk about negative news to family and close friends.

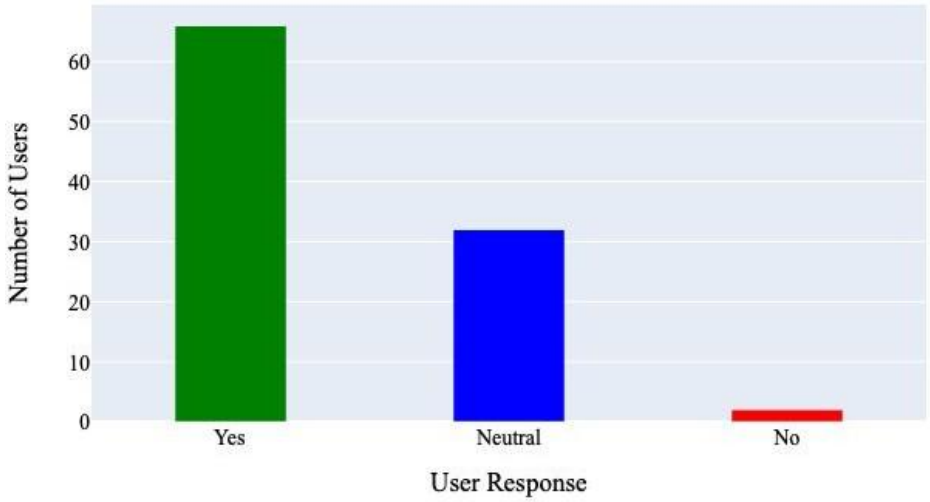

Fig 8. Number of users discussing mostly negative news with close ones

4. *Fig. 9*, shows that 77% of the participants think the computers will play an increasingly important role in journalism.

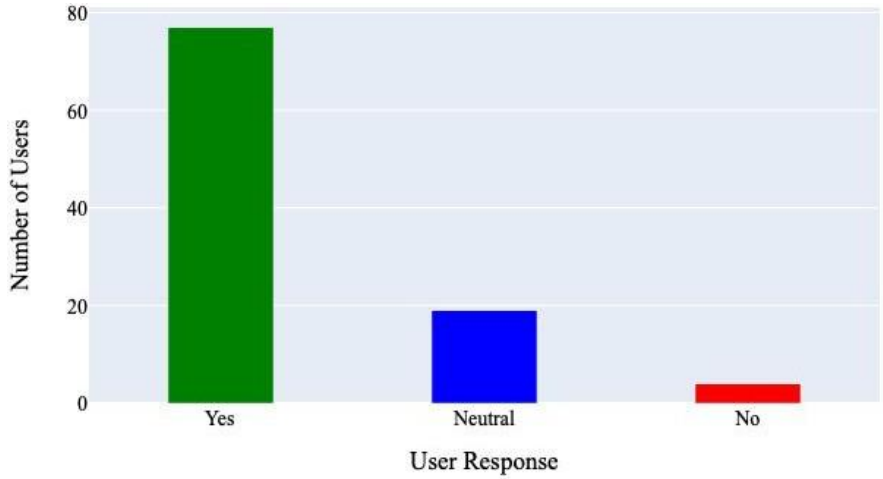

Fig 9. Acceptance of computers in journalism

Future work

The scope of this project in the future can extend into various directions. The major ones of these would be the implementation of a system to evaluate the "interesting-ness" of an article to decide what would be presented to the users before other articles. Another major improvement would be to enable the process of automating the creation of the training dataset of news articles and summaries. This would enable keeping the context of situations more up to date and in tune with time. One of the biggest possibilities of extension that we are looking at is the possibility of creating a system that can train itself along with the collection and labelling of articles. That is, the possibility of automating the process completely and thereby removing the need for any human involvement.

Conclusion

Based on the work done, we see that there is a huge possibility for machine learning to become a regular part of journalism and news media in the future. As machine learning and specifically the field of text processing progresses forward, we may see this change come about very soon. Based on the survey we also understand that the way news is presented to people clearly has an impact on the readers and those who are close to them. Users are also open to change in the field of news media to help tackle issues like sensationalism which should not hold a position in this field.

References

1. Stuart N. Soroka, Good News and Bad News: Asymmetric Responses to Economic Information. In

The Journal of Politics 2006 68:2, 372-385

2. Agarwal, Aashish, Ankita Mandal, Matthias Schaffeld, Fangzheng Ji, Jhiao Zhan, Yiqi Sun and Ahmet Aker. "Good , Neutral or Bad - News Classification." NewsIR@SIGIR (2019)..
3. Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv abs/1907.11692 (2019): n. Pag.
4. Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz and Jamie Brew. "HuggingFace's Transformers: State-of-the-art Natural Language Processing." ArXiv abs/1910.03771 (2019): n. Pag.
5. Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." ArXiv abs/1910.13461 (2019): n.pag.
6. Lin, Chin-Yew. (2004). ROUGE: A Package for Automatic Evaluation of summaries. In Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.
7. Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly D. Voll and Manfred Stede. "Lexicon-Based Methods for Sentiment Analysis." Computational Linguistics 37 (2011): 267-307.
8. Christensen, Heidi & Gotoh, Y. & Kolluru, B.K. & Renals, S.. (2003). Are extractive text summarisation techniques portable to broadcast news?. 489 - 494. 10.1109/ASRU.2003.1318489.
9. Nallapati, Ramesh, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre and Bing Xiang. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond." CoNLL (2016).
10. See, Abigail, Peter J. Liu and Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks." ArXiv abs/1704.04368 (2017): n. Pag.
11. Zhang, Haoyu, Yeyun Gong, Yu Yan, Nan Duan, Jianjun Xu, Ji Wang, Ming Gong and Ming Zhou. "Pretraining-Based Natural Language Generation for Text Summarization." CoNLL (2019).

12. Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." ArXiv abs/1810.04805 (2019): n. Pag.
13. Sun, Chi, Xipeng Qiu, Yige Xu and Xuanjing Huang. "How to Fine-Tune BERT for Text Classification?" CCL (2019).
14. Hermann, Karl Moritz, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman and Phil Blunsom. "Teaching Machines to Read and Comprehend." NIPS (2015).
15. Shi, Tian, Yaser Keneshloo, Naren Ramakrishnan and Chandan K. Reddy. "Neural Abstractive Text Summarization with Sequence-to-Sequence Models." ArXiv abs/1812.02303 (2018): n. pag.