DISSERTATION

Project title: Social Feedback-Enhanced Argument Claim Extraction from News Headlines

COMP3003 - Individual Dissertation

Sea Wei Chuen

Abstract

Nowadays, the way that people obtain their news is changing. The consumption of physical newspapers is gradually decreasing while more and more people obtain their news through the Internet. Not too long ago, people would purchase a newspaper, skim through headlines and read articles that attracted their attention [1]. But nowadays, people often read an article only because it was shared with them on social media or any other internet platforms. Due to the change of how people obtain and read their news from a physical medium to an online medium, this has also allowed social media users themselves to share and create their own news headlines. Paired with the ability for other social media users to comment on the news shared, this has made argumentation mining from social media content to be very rewarding as the large-scale analysis of these social media argumentation can offer very valuable insights.

Table of Contents

1. Introduction and Motivation

Several works have been presented on extracting argument claims from news headlines in social media (e.g. [2], [3]). However, those works did not make use of any social aspect to enhance the argument extraction process. Nowadays, people have the ability to express their thought and opinions on news using various social media services such as leaving comments on news articles. Comments are one of the first kinds of user-generated Web content, and virtually all types of items are commented on, such as texts, images, ideas etc. The very purpose of the comment section of a post is to collect user feedback, but they also provide practical value to visitors as well [4]. This further proves that the comment section in social media is a domain that contains an enormous amount of text data on various subjects and is the place for exchanging opinions. The nature of the comment section in social media is based on debating, which means that it contains plenty of useful information to be extracted [2].

Therefore, the comment section of news can serve as a social aspect. As stated, the importance of the comment section as social feedback towards a news headline cannot be ignored. Therefore, this project explores using the comments as social feedback to enhance argument extraction on news headlines. To accomplish this task, this project uses news headlines related to science obtained from the social media platform Reddit.

Argument mining is a research area in the field of natural language processing. The aim of argument mining is the automatic extraction of natural language corpora from various textual corpora. The main goal is to be able to provide structured data which is machine-passable for computational models of arguments [5]. There are many genres to which argument mining can be applied to such as the qualitative analysis of social media content. It is now a norm where social media users on platforms such as Reddit, Twitter and newspaper blogs can create or share news. These texts are usually short, without standard spelling and with different conventions such as emoticons and hashtags. Due to the unpredictable nature of these user-generated texts, there will be some unique challenges faced when argument mining on social media data. This gives rise to the need for defining methods to extract accurate argument claims from headlines [3].

An argument has 2 parts, and the argument must have 1 central claim. In an argument, it is the claim that is the most important component. It is a controversial statement that should not be accepted by the reader without additional support, whereas the premise underpins the validity of the claim. The premise is a reason given by the person making the argument for persuading the listener or reader of the claim [6]. This dissertation focuses on extracting only the claim as it is the central component of an argument.

Headlines play a crucial role in attracting the attention of the audience to online content. In this digital age, there is a rapidly increasing gap between limited attention and limitless media that makes it difficult for anything to attract an audience. It is found that most social media users spend more time browsing and scanning for specific keywords as well. There is less time spent by users on concentrated reading, and sustained attention is decreasing [7]. Therefore, being able to extract the main argument claim from news headlines, this allows social media platforms to be able to summarise news headlines and create more personalised recommendations. Furthermore, argument extraction can be a useful tool for lawmakers and researchers in social sciences.

2. Description of the work

The aim of this project is to design a method for extracting the argument claim from news headlines making use of a social aspect such as user comments to enhance the argument claim extracted. The

methods designed are from the domains of artificial intelligence and machine learning in combination with natural language processing techniques. The end goal would be the development of a natural language method that can perform the aim of this project with reasonably good accuracy and low error. The natural language processing method should also be able to perform reasonably well to other news headline datasets with acceptable computational cost.

Objective:

- To produce a full literature search and literature review of existing related works.
- To efficiently and accurately extract the argument claims through the use of natural language processing methods.
- To determine an appropriate method of adopting a social aspect to enhance argument claim extraction.
- To evaluate the method using suitable statistical techniques.

3. Related work

3.1 Argument extraction

3.1.1 Argument Extraction from News

In [2], the authors have done argument extraction on blogs and news. They have used a supervised approach is based on Conditional Random Fields (CRF) for argument extraction. The authors try to detect the segments in sentences that represent the main claims and premises of the sentence. The approach proposed by the authors exploits features such as distributed representations of words, words, part-of-speech tags, and small lists of language-dependent cue words.
The results that they have achieved using the proposed approach to predict the claim of the sentence were a precision score of 38.72%, a recall of 27.60% as well as an F1 score of 32.21%.

3.1.2 Automatic Extraction of News Values from Headline Text

In [7], the authors present a new perspective on digital content processing by creating an automatic news headline processor which extracts news values to help in the customised selection of digital content to users. The authors have used notable natural language processing methods such as sentiment analysis, wikification, and language modelling. They have also combined those methods with ai methods like burst detection algorithm. Using the mentioned techniques, the authors have categorised the news headlines into several categories.

3.1.3 Argued opinion extraction from festivals and cultural events on Twitter

In [9], the authors attempt to perform a supervised argument extraction from Twitter data. They have done several steps of data processing before evaluation. The first step was to convert the Twitter data into vectors. They have made use of the *word2vec* tool to achieve that. Next, they have extracted the tweets which contain an argument. This was done using machine learning binary classifiers like Random Forest, Logistic Regression, Support Vector Machines and Naïve Bayes in a supervised setting. Finally, they have used CRFs to extract the claims of the tweets.

The authors in [9] have achieved a precision of 87.00%, a recall of 71.00% as well as an F1 score of 78.00%

3.1.4 Context-Independent Claim Detection for Argument Mining

In [10], the authors attempt to detect the presence of a claim from sentences in text documents. They have proposed a methodology that uses a Scalar Vector Machine (SVM) based Partial Tree Kernels.

The authors have compared their methodology with another popular claim detection methods bag of words. It is found that their SVM based Partial Tree Kernel method produces a significant improvement in precision, recall and f1 score (improvement of +1.60%, +7.00% and +4.6%, respectively).

3.1.5   A Question Answering-Based Framework for One-Step Event Argument Extraction

In [11], the authors have done a one-step event argument extraction using a question and answer based framework for a single-step event argument extraction. The proposed method provides an easy way to solve argument role classification and argument candidate extraction at the same time. The proposed method uses a pre-trained model from BERT.

The results obtained were a precision of 53.30%, recall of 48.20% and an F1 score of 50.60%.

3.2 Sentiment analysis

3.2.1   Sentiment analysis on social media for stock movement prediction

In [12], the authors have used sentiment analysis on social media data together with historical stock market data to enhance the prediction of the movement of the stock. They have compared different sentiment analysis methods such as Latent Dirichlet Allocation (LDA), joint sentiment/topic model (JST) and aspect-based sentiment analysis.

It was found that aspect-based sentiment analysis managed to correctly predict the stock market movement with the highest accuracy achieving an average accuracy of 54% – 56% when tested with a different threshold.

3.2.2   Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach

In [13], the authors make use of the hashtags used in tweets to further improve the results of the sentiment analysis. The authors have used a two-stage SVM classifier to predict the sentiment polarity of a specific tweet.

The results obtained was good, with an accuracy of 88.96%, precision of 90.49% and an F1 score of 92.60%.

3.2.3   Hybrid sentiment classification on twitter aspect-based sentiment analysis

This paper [14] proposes the use of aspect-based sentiment analysis on Twitter data. The authors have done an aspect-based feature extraction using association rule mining and the Stanford dependency parser method. This is followed by the usage of the principal component analysis (PCA) algorithm for feature selection. Finally, they have used an SVM classifier for sentiment classification.

They have tested the proposed method with several Twitter datasets and have gained an average precision, recall and f1 score of 70% - 90% for all datasets.

3.2.4   Twitter as a Corpus for Sentiment Analysis and Opinion Mining

This paper [15] presents a way for the automatic collection of Twitter data to train a sentiment classifier. The authors proposed a method of sentiment analysis on Twitter data using a multinomial Naïve Bayes sentiment classifier. Before using the Naïve Bayes classifier, the authors have used the N-grams and part of speech (POS) tags for feature extraction. They have experimented on unigrams, bigrams and trigrams, with unigrams yielding the best results with an accuracy of around 65.00% and an F1 score of 60.20%.

4. Methodology

4.1 Data

Due to lack of consistently annotated news headline data, which has a social aspect to it. I decided to scrape my own news headline data from the Internet along with their comments.

Reddit is a very popular news aggregation website that was created in America. Users of Reddit could post and be involved in discussions about these posts. The Reddit platform is extremely diverse, with over 300,000 subreddits, and each subreddit represents a specific topic. These threaded discussions provide a large news corpus paired along with the social responses to them. The structure of conversations in Reddit are threaded. Each submission (Reddit post) may have a number of top-level comments, and every comment can be commented on in response as well [16].

With about 430 million average monthly active users and also known to be the "face of the internet" by many, Reddit has proven itself to be one of the most popular internet forums. Reddit is a social network that contains hundreds of groups known as "subreddits". A Reddit user can join a subreddit that is topic-oriented of their interest and create a post in the subreddit. Other users can then comment on the subreddit and even upvote or downvote the subreddit, which is essentially the same as a like and dislike. Due to the large amounts of subreddits and comments available on Reddit, it has become a valuable source for social media data collection and analysis. Compared to other social media platforms, Reddit is considerably more open to data collection, making it the subject of numerous scientific studies. Due to all the reasons stated above, the news headlines and social feedback data used for this project will be from Reddit.

To extract the submission and comments data from Reddit, the Pushshift application programming interface (API) was used. Pushshift is a great tool for scrapping Reddit data as it makes it easy to query and obtain past Reddit data. Pushshift can return subreddit submissions as well as the associated comments posted in a specified period of time. Most importantly, Pushshift also has a significantly larger single query size limit than the official Reddit api.

As Reddit submissions and comments contain the usernames of Reddit users, anonymisation steps have been taken to ensure the anonymity of the Reddit users. This was done using the Pushshift api parameters. Table 1 shows the data that was obtained during the query after anonymisation.

sub_id	headlines	publish_date	num_of_comments

Table 1: Header of data obtained after anonymisation

Since this project uses news headline data, the "r/science" subreddit will be investigated. The submissions and comment that were posted between 1st October 2020 to 13th November 2020, there are a total of around 200 submissions obtained. The main reason why the "r/science" subreddit was used is that this subreddit is heavily moderated. Every submission had to be peer-reviewed research, and most importantly, the subreddit has strict rules against misleading clickbait titles, joke comments and fake news.

4.2 Data labelling

When labelling data, longer news headlines have been manually chosen for the argument extraction task as longer headlines are more likely to contain the central argument claim as well as the supporting premise, which would be able to test the proposed method to extract the argument claim more accurately. The news headlines were manually labelled by 2 human annotators by choosing the sentence where they felt was the central argument claim. During the labelling process, there were some disagreements between the 2 human annotators on certain headlines, but a consensus was reached after some discussion. Table 2 shows the data after submission selection and labelling.

headlines	labelled_argument
Greenland could lose more ice this century than it has in 12,000 years- "The paper is also an answer to those who dismiss the ongoing effects of climate change with 'the earth has always changed'-and the answer is, 'not at this pace'," Scambos says.	Greenland could lose more ice this century than it has in 12,000 years
COVID-19 lockdown has changed how we use recreational drugs. Increased use of alcohol due to boredom and free time, decreased use of stimulants because clubs and bars are closed	COVID-19 lockdown has changed how we use recreational drugs.

Table 2: Labelled data

4.3 Data cleaning

Since the data was obtained through a social media platform, there were quite a number of unusable submissions and comments. The submissions were filtered using a few characteristics, such as removing deleted submissions (due to deactivated accounts) and removing submissions with less than 100 comments. At the same time, the comments that were used are only top-level comments of the submission. After removing the unusable submissions and comments, the Jason data obtained from the Pushshift api was exported into a more readable Microsoft excel file format.

4.4 Latent Dirichlet Allocation (LDA) [17] for Topic Modelling

This is a brief explanation of the words that form the name of this topic modelling technique. 'Latent' refers to a hidden or concealed property of the document which the LDA model tries to find. 'Dirichlet' is the assumption of LDA that the distribution of words in topics and the distribution of

topics in a document are both Dirichlet distributions. Lastly, 'Allocation' refers to how topics in the document are distributed.

LDA is a generative probabilistic model of a collection of documents made up of words. LDA assumes that a document is made up of words, and these words all belong to a topic. Subsequently, LDA generates several topics in the document and assigns words contained in the document to the generated topics. When each word is assigned to a topic, the order in which the words occurs, along with all syntactic information, is lost. Therefore, LDA is considered to be a bag-of-words model. It is also important to note that the topics produced will not be named, and it is often necessary for a human annotator to look at the words assigned to the topics to determine if the topics have been modelled correctly and label the topics.

The LDA algorithm starts by randomly assigning words to a topic, and then it uses Gibbs sampling [19] iteratively to improve the assignment words to topics. The number of topics that the algorithm will generate, as well as the number of iterations, can be configured using the NLP library used.

5. Design and Implementation

From the start of this project until the end, Python was chosen as the primary programming language of the project. Python was chosen because of the extensive libraries available for NLP related tasks as well as its simple to understand syntax. Although the execution speed of Python is slower than the other famous programming languages such as Java or C++, the pros outweigh the cons by a lot.

Currently, the system design consists of 4 main components to argument claim extraction. The first component is the data collection system, which extracts Reddit data using the Pushshift api. The second component is the pre-processing system, which performs stop words removal, lemmatisation, and tokenisation, the third is the argument extraction section using LDA topic modelling, and the fourth component is the evaluation component where the LDA topic words are compared with the labelled data and evaluated using the bleu score.

5.1 Data Collection

A lot of work and effort has been put into data collection due to the lack of a proper news headline dataset that has comment data included. It was decided during the design process that the Reddit data to be used are from 1st October 2020 to 31st October 2020. To accomplish the data collection task, server requests have been made to the pushshift api iteratively. In every iteration, the headline the submissions are extracted and placed into a pandas dataframe, which is essentially a data structure that makes data manipulation easier. The headline data is then exported into a Microsoft Excel file. The same process was repeated for comment data.

5.2 Data Pre-Processing

Like in most NLP tasks, there are some simple pre-processing techniques used before performing argument extraction. Firstly, the headlines and comments data has been tokenised. Tokenisation is the task of splitting up a text string into small units, which are called tokens. The type of tokenisation that was used was white space tokenisation, and digits were removed as well. Subsequently, stop words have been removed using the "stop-words" python package. Stop words are common words that appear commonly in the English language. Removing stop words can help improve NLP task by removing noisy words. The final pre-processing step was to lemmatise the data, lemmatising cut words into their root form. The advantage of lemmatising is that similar words in different forms will become the same word, e.g. (foots and feet both become foot).

The lemmatisation and tokenisation tasks were carried out using the Natural Language Toolkit (NLTK) python library. Whereas stop words removal was done according to the "stop-words" list of stop words available in the pip python package manager. Table 3 and 4 show 3 out of 206 submissions and comment data before and after pre-processing.

headline	comments
Greenland could lose more ice this century than it has in 12,000 years-"The paper is also an answer to those who dismiss the ongoing effects of climate change with 'the earth has always changed'-and the answer is, 'not at this pace'," Scambos says.	The people still questioning climate change will be the same people watching their beachfront property float away in a few years they can't fathom something till it alters their life. The problem is that most people will not bother reading the paper or even listening to scientists. People still believe that the earth is only a few thousand years old. Some people still believe the earth is flat; that Bill Gates created COVID-19; that vaccines cause autism; I know of some these people personally, they didnt do very well in high school...you would think that these people do not get very far in life, but I know of three politicians, and a few real estate agents.
Lesbians, gays, bisexuals experience migraines at a rate 58% higher than heterosexual or "mostly straight" people. 1 in 6 adults will have migraine headaches in their lifetime, but among LGB people, that figure is 1 in 3. Researchers speculate that added stress and discrimination may be responsible.	Added stress and discrimination? Not to offend anyone, but wouldn't that also affect certain races then? Link to abstract:- [Disparities Across Sexual Orientation in Migraine Among US Adults](https://jamanetwork.com/journals/jamaneurology/article-abstract/2771029) This seems like causation without correlation to me. Damn, as if being discriminated against isn't bad enough. Could this be something to do with different body chemistry I wonder?
The prevalence of dementia in countries where more than one language is spoken is 50% lower than in those regions where the population uses only one language to communicate. Active bilingualism is an important predictor of delay in the onset of symptoms of mild cognitive impairment.	Correlation does not imply causation. Instead of assuming that the language proficiency is protecting them from dementia, both of these things may have been caused by a third factor. For example, the more intelligent and better educated you are, the more you are likely to be proficient in both languages. Oh man the US is so fucked. But this actually explains a lot. What if I'm happier actively ignoring instead of participating? Good thing i know java, javascript, and Python Explains America.

Table 3: submissions and comments data before pre-processing

headline_preprocessed	headline_comment_preprocessed

[['greenland', 'lose', 'ice', 'century', 'year', 'paper', 'also', 'answer', 'dismiss', 'ongoing', 'effect', 'climate', 'change', 'earth', 'always', 'changed', 'answer', 'pace', 'scambos', 'say']]	[['greenland', 'lose', 'ice', 'century', 'year', 'paper', 'also', 'answer', 'dismiss', 'ongoing', 'effect', 'climate', 'change', 'earth', 'always', 'changed', 'answer', 'pace', 'scambos', 'say', 'climate', 'change', 'paper', 'earth', 'earth', 'paper', 'paper', 'climate', 'change', 'also', 'greenland', 'earth', 'climate', 'change', 'always', 'always', 'earth', 'climate', 'change', 'lose', 'greenland', 'earth', 'always', 'change', 'earth', 'always', 'also', 'climate', 'ice', 'earth', 'ice', 'always', 'earth', 'change', 'also', 'change', 'ice', 'ice', 'ice', 'paper', 'paper', 'also', 'answer', 'dismiss', 'ongoing', 'effect', 'climate', 'paper', 'also', 'answer', 'dismiss', 'ongoing', 'effect', 'climate', 'change', 'earth', 'always', 'answer', 'scambos', 'say', 'ice', 'ice', 'ice', 'ice', 'always', 'ice', 'ice', 'climate', 'change', 'climate', 'change', 'also', 'ice', 'change', 'greenland', 'always', 'earth', 'climate', 'greenland', 'climate', 'change', 'ice', 'climate', 'change', 'ice', 'ice', 'greenland', 'climate', 'change', 'always', 'change', 'climate', 'effect', 'ice']]
[['lesbian', 'gay', 'bisexual', 'experience', 'migraine', 'rate', 'higher', 'heterosexual', 'mostly', 'straight', 'people', 'adult', 'migraine', 'headache', 'lifetime', 'among', 'lgb', 'people', 'figure', 'researcher', 'speculate', 'added', 'stress', 'discrimination', 'may', 'responsible']]	[['lesbian', 'gay', 'bisexual', 'experience', 'migraine', 'rate', 'higher', 'heterosexual', 'mostly', 'straight', 'people', 'adult', 'migraine', 'headache', 'lifetime', 'among', 'lgb', 'people', 'figure', 'researcher', 'speculate', 'added', 'stress', 'discrimination', 'may', 'responsible', 'stress', 'experience', 'experience', 'experience', 'stress', 'stress', 'experience', 'migraine', 'migraine', 'higher', 'rate', 'migraine', 'migraine', 'discrimination', 'migraine', 'headache', 'stress', 'stress', 'migraine', 'migraine', 'stress', 'stress', 'lgb', 'stress', 'stress', 'stress', 'rate', 'heterosexual', 'migraine', 'migraine', 'migraine', 'migraine', 'migraine', 'migraine', 'stress', 'migraine']]
[['prevalence', 'dementia', 'country', 'one', 'language', 'spoken', 'lower', 'region', 'population', 'us', 'one', 'language', 'communicate', 'active', 'bilingualism', 'important', 'predictor', 'delay', 'onset', 'symptom', 'mild', 'cognitive', 'impairment']]	[['prevalence', 'dementia', 'country', 'one', 'language', 'spoken', 'lower', 'region', 'population', 'us', 'one', 'language', 'communicate', 'active', 'bilingualism', 'important', 'predictor', 'delay', 'onset', 'symptom', 'mild', 'cognitive', 'impairment', 'one', 'language', 'language', 'dementia', 'language', 'country', 'language', 'spoken', 'population', 'bilingualism', 'language', 'spoken', 'language', 'population', 'one', 'language', 'dementia', 'one', 'cognitive', 'impairment', 'country', 'country', 'one', 'language', 'one', 'delay', 'onset', 'mild', 'cognitive', 'language', 'cognitive', 'one', 'one', 'one', 'language', 'prevalence', 'dementia', 'dementia', 'language', 'bilingualism', 'cognitive', 'dementia', 'one', 'language', 'bilingualism', 'dementia', 'one', 'bilingualism', 'language', 'dementia', 'country', 'important', 'one', 'one', 'country', 'dementia', 'region', 'country', 'one', 'language', 'us', 'language', 'dementia', 'lower', 'country', 'bilingualism', 'one', 'one', 'language', 'cognitive', 'dementia', 'one', 'dementia']]

Table 4: submissions and comments data after pre-processing

5.3 Argument Extraction with LDA

Before LDA is performed, the comment data is processed further by removing all words that do not appear in the headline title. After this step, the comment data obtained will be roughly indicative of which parts of the headline title are talked about by reddit users most often.

LDA is a generative probabilistic model. When given a single document as input, it will assign words to topics along with the calculated probability of the word being in that topic. Therefore, there are 2 important LDA parameters which need to be tuned, which are the number of topics as well as the top n number of words which contributes to the topic. During this project, the assumption that each news headline contains only 1 argument has been made. Therefore, the number of topics of 1 has been chosen. Tuning the top n number of topic words is important because that is essentially the length of the argument claim extracted from news headlines. Ideally, there would be a method which given the headline, the method can effectively "guess" the length of the argument claim and use that length as n. Therefore, a series of experiments have been conducted to determine the best n to be used for argument extraction. The top n number of words to be used will be investigated in the evaluation section. Table 5 shows the output after performing LDA on headlines and the social feedback enhanced headline (headline combined with the comments) where n is set to be half of the length of the headline string.

headlines	labelled_argument	headline_lda	social_aspect_lda
Lesbians, gays, bisexuals experience migraines at a rate 58% higher than heterosexual or "mostly straight" people. 1 in 6 adults will have migraine headaches in their lifetime, but among LGB people, that figure is 1 in 3. Researchers speculate that added stress and discrimination may be responsible.	Lesbians, gays, bisexuals experience migraines at a rate 58% higher than heterosexual or "mostly straight" people.	people migraine added speculate responsible researcher rate mostly may lifetime lgb lesbian higher heterosexual headache gay figure experience discrimination bisexual among adult straight stress	migraine stress experience rate discrimination people headache heterosexual higher lgb speculate responsible researcher mostly added lifetime straight gay figure bisexual among adult may lesbian

The prevalence of dementia in countries where more than one language is spoken is 50% lower than in those regions where the population uses only one language to communicate. Active bilingualism is an important predictor of delay in the onset of symptoms of mild cognitive impairment.	Active bilingualism is an important predictor of delay in the onset of symptoms of mild cognitive impairment.	one language active spoken region prevalence predictor population onset mild lower important impairment dementia delay country communicate cognitive bilingualism symptom us	language one dementia country bilingualism cognitive spoken population lower region prevalence onset us important impairment delay mild symptom predictor communicate active
Daily alcohol intake triggers aberrant synaptic pruning leading to synapse loss and anxiety-like behavior. A chronic binge drinking protocol resulted in depressed neurotransmission and increased anxiety-like behaviors in mice by activating microglia that destroy neuronal connections.	Daily alcohol intake triggers aberrant synaptic pruning leading to synapse loss and anxiety-like behavior.	like anxiety behavior leading chronic binge alcohol depressed connection aberrant loss microglia mouse neuronal neurotransmission protocol pruning resulted synapse synaptic intake daily increased	drinking alcohol binge mouse anxiety behavior like intake chronic daily depressed increased protocol loss destroy aberrant neurotransmission resulted activating synapse synaptic pruning connection

Table 5: LDA on headlines as well as social feedback enhanced headline

5.4 Argument Extraction with LDA

As mentioned before in the explanation of LDA, it is a bag of words model. Therefore, as seen in table 5 the output does not preserve the sequence of words. Due to this limitation, the evaluation method used does not penalise the sequence of the output.

5.5 Baseline argument extraction

To set a benchmark for the performance of argument extraction without social feedback enhancement, a baseline argument extraction has been carried out. The baseline method used is to perform LDA on only the headline. This means that the document (headline) used as input for LDA will be of a smaller size and typically relies on how often a word appears in the headline itself.

5.6 Proposed social feedback enhanced argument extraction

The proposed method to achieve this argument extraction task is to combine the pre-processed headline and comments first before performing LDA. Based on the assumption that the comments all discuss the main argument claim of the headline, this should cause the expected argument words to have a higher word count in the new combined document (headline and comments). Therefore, increasing the contribution of the word to the topic as there is only 1 topic generated by LDA at the

current setting, causing LDA to pick the expected argument words with a higher probability. This is the hypothesis that was made when doing this project.
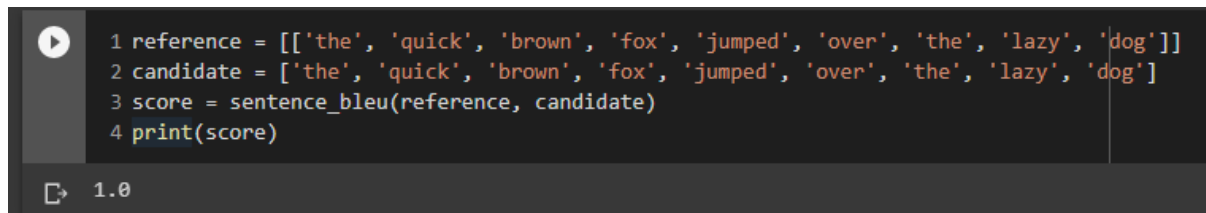
6. Evaluation

6.1 BLEU metric

The primary programming task for a BLEU implementor is to compare *the n*-grams of the candidate with the *n*-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation is. For simplicity, we first focus on computing unigram matches.

The counting of matching n-grams is modified to ensure that it takes the occurrence of the words in the reference text into account, not rewarding a candidate translation that generates an abundance of reasonable words. This is referred to in the paper as modified n-gram precision.

The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. It is important to note that the more reference translations per sentence there are, the higher the score

is. Thus, one must be cautious in making even "rough" comparisons on evaluations with different numbers of reference translations.

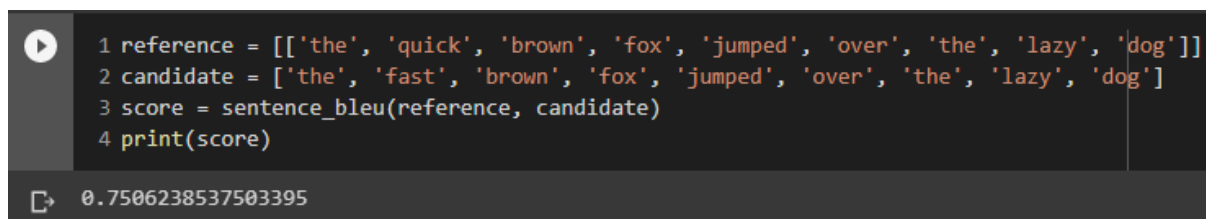This is an example of a candidate who gives a perfect BLEU score of 1.

```
1 reference = [['the', 'quick', 'brown', 'fox', 'jumped', 'over', 'the', 'lazy', 'dog']]
2 candidate = ['the', 'quick', 'brown', 'fox', 'jumped', 'over', 'the', 'lazy', 'dog']
3 score = sentence_bleu(reference, candidate)
4 print(score)

1.0
```

Diagram 6: Perfect BLEU score of 1

Whereas a slight change in the candidate causes a slight drop in score.

```
1 reference = [['the', 'quick', 'brown', 'fox', 'jumped', 'over', 'the', 'lazy', 'dog']]
2 candidate = ['the', 'fast', 'brown', 'fox', 'jumped', 'over', 'the', 'lazy', 'dog']
3 score = sentence_bleu(reference, candidate)
4 print(score)

0.7506238537503395
```

Diagram 7: BLEU score of 0.75 when "fast" is interchanged with "quick"

6.2 LDA results

Several experiments have been carried out using different values of n to investigate the effect of the top n number of topic words used on the average bleu score of the entire dataset as well as individual headlines. Table 8 shows the results obtained.

top n words of LDA	Average_baseline_bleu_score	Average_social_aspect_bleu_score
n = half of headline + 7	0.66	0.63
n = half of headline + 5	0.67	0.63

n = half of headline + 3	0.66	0.63
n = half of headline	0.66	0.65
n = half of headline - 3	0.65	0.63
n = half of headline - 5	0.65	0.63
n = half of headline - 7	0.66	0.65
n = half of headline - 10	0.60	0.60
n = half of headline - 11	0.57	0.56
n = half of headline - 12	0.53	0.53
n = half of headline - 13	0.48	0.49

Table 8: several values of n have been used

Based on the results obtained, when n is of a high value (n >= half of headline – 7) the average BLEU scores of both methods are relatively higher than when a smaller n is used. At high values of n, the baseline method performs better than the social feedback enhanced method. Whereas, at lower values of n (n <= half of headline – 10), the social feedback enhanced method performs almost as well as the baseline method.
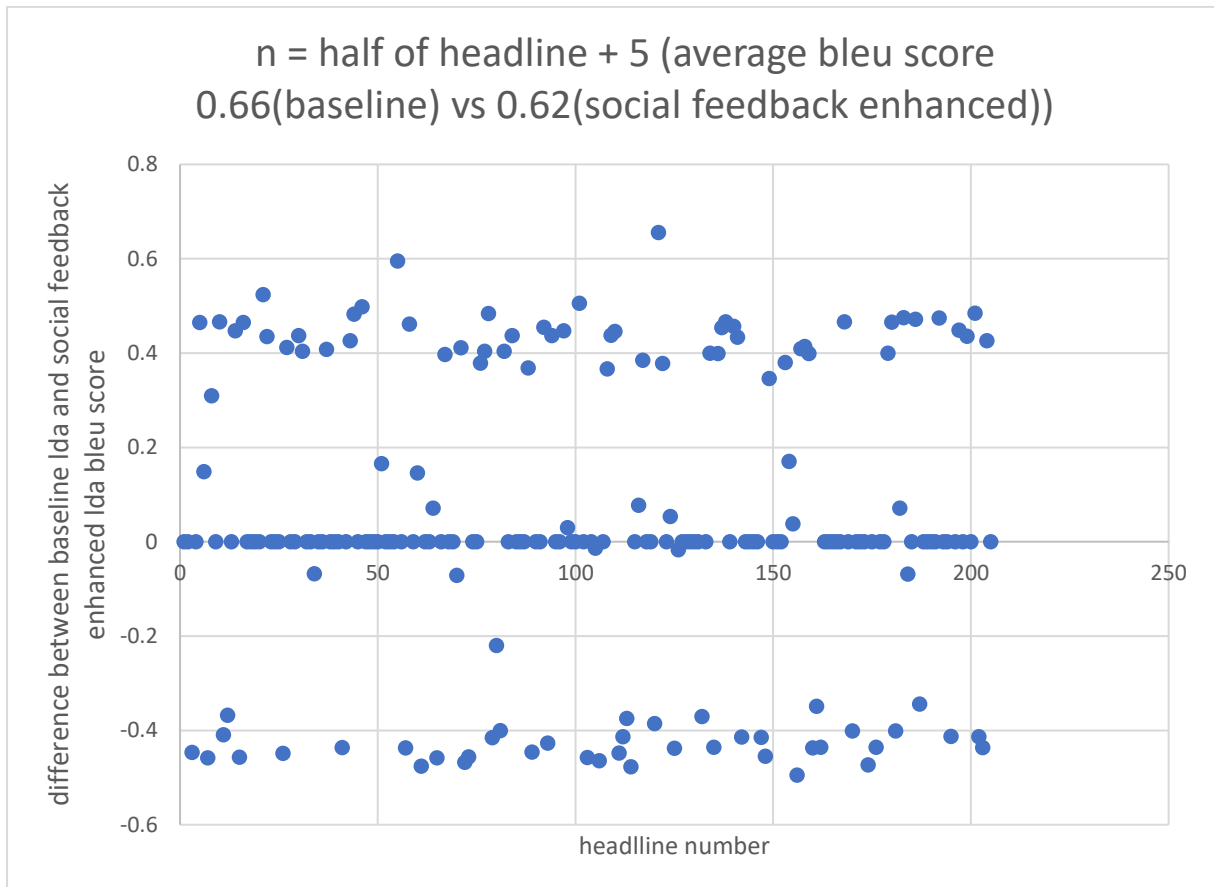

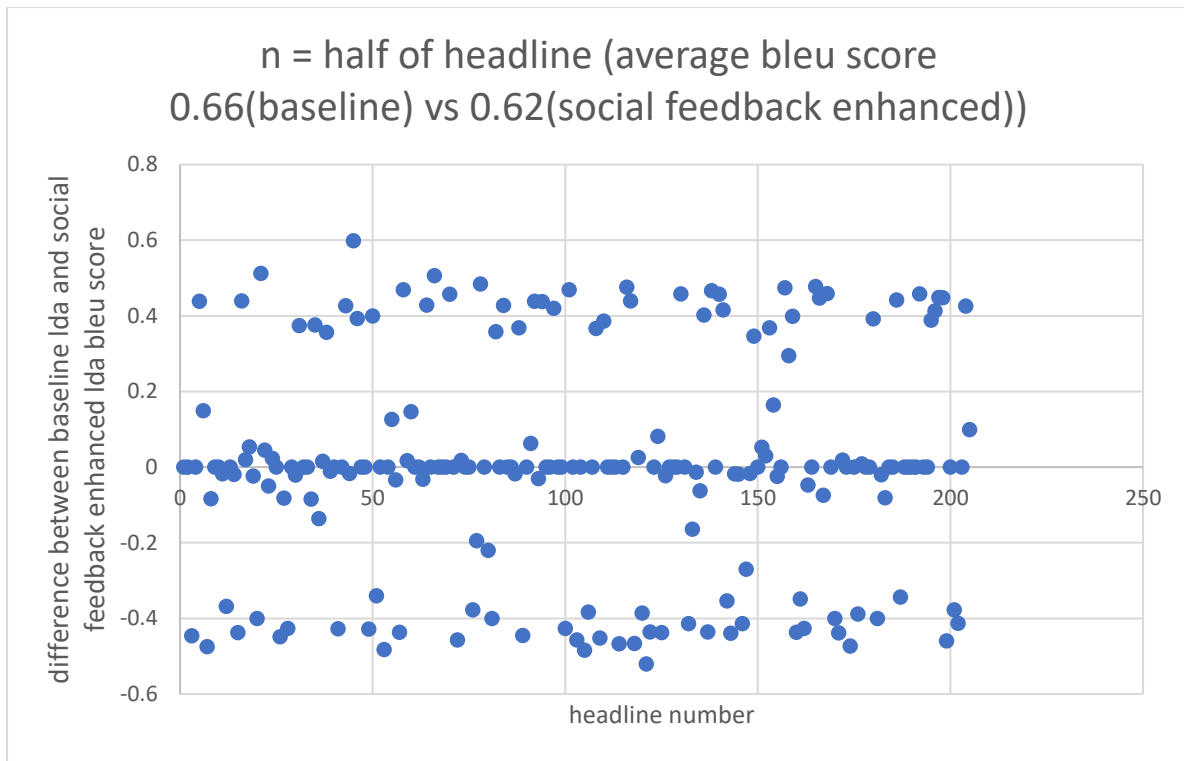Figure 9: difference of bleu scores when n = half of headline + 5

Figure 10: difference of bleu scores when n = half of headline
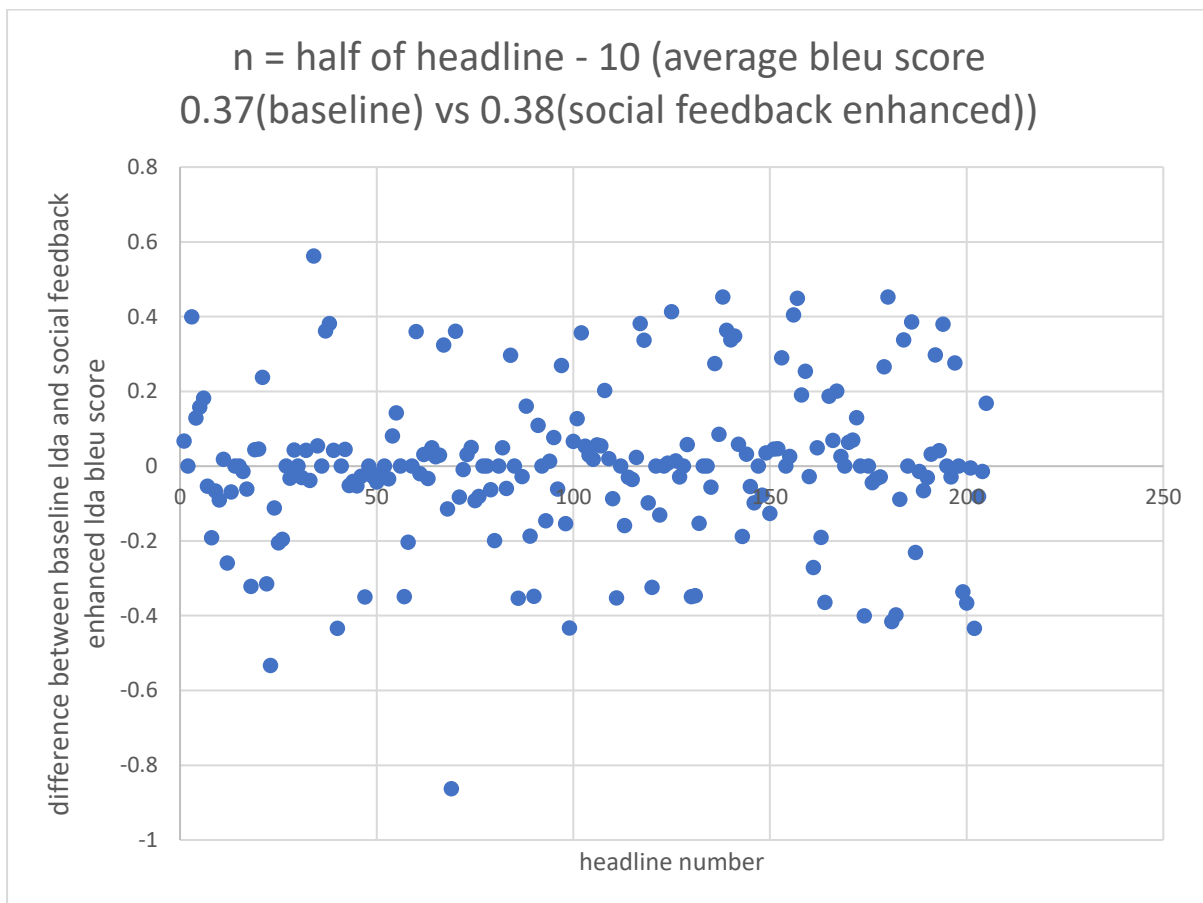

Figure 11: difference of bleu scores when n = half of headline - 10

The graphs in figure 9 to figure 11 above shows that for most headlines, the social feedback enhanced method and baseline methods perform equally well. However, there are some headlines where the socially enhanced method performs better and vice versa. It can also be seen that as n is decreased, the difference in the 2 LDA methods is more obvious. This is shown when most headlines have a non-zero difference between baseline method LDA bleu scores and social feedback enhanced LDA bleu scores when a low value of n is used. The opposite is true when the value of n is increased. This shows that enhancing the news headlines with social feedback does indeed have an effect when using LDA for argument extraction. However, it does not improve the average bleu score obtained on this specific dataset.

6.3 LDA result analysis

For this result analysis, only the top n topic word of n = half of headline – 10 will be considered as it has the largest overall difference between baseline LDA and social feedback enhanced LDA bleu score. This implies that the impacts of social feedback enhanced LDA can be observed more clearly.

Table 12 shows the LDA output of when the social feedback enhanced method performed better than the baseline method. Social feedback enhanced headlines are found to perform better on labelled headlines which are of shorter length. Along with the comments mentioning the argument claim more frequently, this has caused LDA to be able to detect the argument words more accurately.

Table 13 shows the LDA output of when the social feedback enhanced method performed worse than the baseline method. Social feedback enhanced headlines are found to perform worse on labelled headlines that are of longer length. The comments for those headlines also seem to mention other aspects of the headlines more frequently, causing the extracted argument to have a lower bleu score.

Table 12: LDA output of when social feedback enhanced headlines perform better.

headlines	labelled_argument	headline_lda	social_aspect_lda	baseline_bleu_score	social_aspect_bleu_score
6% of adults hospitalised with COVID-19 in the US were healthcare workers. Among them, 36% were in nursing-related occupations. Approximately 28% of these patients were admitted to an ICU, 16% required invasive mechanical ventilation, and 4% died.	['adult', 'hospitalized', 'covid', 'healthcare', 'worker']	mechanical icu ventilation required related patient occupation nursing	healthcare hospitalised adult covid patient worker admitted died	0	0.863340021
Immune cell activation in severe COVID-19 resembles lupus. This may explain why some people infected with SARS-CoV-2 produce abundant antibodies against the virus, yet experience poor outcomes.	['immune', 'cell', 'activation', 'severe', 'covid', 'resembles', 'lupus']	produce sars people poor	covid lupus infected produce	0	0.533085912
Older adults with severe apathy, or lack of interest in usual activities, may have a greater chance of developing dementia than people with few symptoms of apathy, according to a study. Apathy may be a very early sign of dementia and it can be evaluated with a brief questionnaire.	['apathy', 'may', 'early', 'sign', 'dementia', 'evaluated', 'brief', 'questionnaire']	apathy may dementia greater symptom study sign severe questionnaire people older according lack developing	dementia people apathy interest early may severe brief chance adult sign developing symptom questionnaire	0.392814651	0.826516818
Vitamin D supplementation for 12 months appears to improve cognitive function through reducing oxidative stress regulated by increased telomere length (TL) in order adults with mild cognitive impairment (MCI). Vitamin D may be a promising public health strategy to prevent cognitive decline.	['vitamin', 'may', 'promising', 'public', 'health', 'strategy', 'prevent', 'cognitive', 'decline']	cognitive vitamin regulated stress promising public reducing oxidative tl telomere strategy	vitamin cognitive order increased adult health decline prevent strategy public improve	0.441162936	0.873935133

headlines	labelled_argument	headline_lda	social_aspect_lda	baseline_bleu_score	social_aspect_bleu_score
For the first time scientists have made a discovery of hallucinogenic drug remains at the site of an early California cave painting. These drug remnants were discovered in the ceiling of Pinwheel Cave and could change scientists' understanding of how the community used that space -- and the drugs.	['first', 'time', 'scientist', 'made', 'discovery', 'hallucinogenic', 'drug', 'remains', 'site', 'early', 'california', 'cave', 'painting']	drug scientist cave pinwheel hallucinogenic understanding time space site remnant remains california painting discovered	cave drug used time scientist pinwheel hallucinogenic community change remains understanding ceiling space site	0.880111737	0.427287006
New research could help millions who suffer from 'ringing in the ears': Researchers show that combining sound and electrical stimulation of the tongue can significantly reduce tinnitus, commonly described as "ringing in the ears"; therapeutic effects can sustain for up to 12 months post-treatment	['researcher', 'show', 'combining', 'sound', 'electrical', 'stimulation', 'tongue', 'significantly', 'reduce', 'tinnitus']	ringing ear research tongue tinnitus therapeutic sustain suffer stimulation sound significantly show	tongue sound help tinnitus ringing research electrical stimulation month suffer combining ear	0.82423675	0.442850014
Scientists from the Tokyo University of Science have made a breakthrough in the development of potential drugs that can kill cancer cells. They have discovered a method of synthesising organic compounds that are four times more fatal to cancer cells and leave non-cancerous cells unharmed.	['scientist', 'tokyo', 'university', 'science', 'made', 'breakthrough', 'development', 'potential', 'drug', 'kill', 'cancer', 'cell']	cell cancer non leave unharmed tokyo time synthesizing scientist science potential organic	cancer cell breakthrough time fatal tokyo non leave cancerous university drug compound	0.82423675	0.442850014

Table 13: LDA output of when social feedback enhanced headlines perform better

7 Summary and Reflections

Based on the results of the experiments, it is found that performing LDA on social feedback enhanced headlines with the pre-processing steps taken do not improve the overall bleu score of the dataset obtained but, depending on the parameter n used, can be equally as good or slightly worse than the baseline method. However, there are some visible patterns on what the social feedback enhanced approach can achieve. It is shown that the proposed method generally performs better when the argument claim of the headline is short relative to the length of the headline (significantly less than half the length of the headline). Whereas the overall topics mentioned in the comments can heavily skew the arguments extracted.

This project explores using LDA to extract social feedback enhanced news headlines from Reddit. Throughout the course of the project:

- A full literature search and literature review of existing related works have been done.
- An argument extraction method that implements social feedback which can extract arguments in a short time and with low computational costs has been developed
- A decent method for implementing a social aspect to enhance argument claim extraction has been explored (taking words in comments to boost the weights of the matching words in headlines)
- The method has been evaluated using a suitable technique (BLEU score)

Overall, the project was a success as the main goals have been accomplished. The project has explored the method of argument extraction that it was supposed to.

7.Future Works

There are many possible future works that can be done on many aspects of the project. The obvious first steps include, would be the creation of a news headline labelled dataset, which includes social feedback such as comments for the purpose of argument extraction. Ideally, all the comments would be discussing the main argument claim of the headline as well. Additionally, it could also be worth exploring different pre-processing techniques. Effective data cleaning and pre-processing is crucial to NLP tasks such as argument extraction. Other areas for possible improvements would be exploring the usage of other argument extraction methods which can preserve the sequence of words. Currently, LDA can only output the top topic words without any information of the sequence.

7.1 Project management

This project was approached in a waterfall manner where tasks were completed sequentially according to the Gantt chart created during the project proposal. The project has a total of 4 milestones: documentation and planning, initial architecture development, implementation, and dissertation writing. The first 2 milestones (documentation and planning) were completed as planned during semester 1. Whereas the last 2 milestones did not go as smoothly as planned in semester 2. This was due to the heavier workload in terms as I was taking 70/120 credits in semester 2. The heavier workload has caused some delays in the delivery of certain project tasks. However, most of the catch- up work of the project has been done during the Easter holidays. Overall, semester 1 went as planned, and semester 2 had some hiccups along the way. However, the project quality did not suffer because of that.

Table 14: Project Work Plan

TASK	DURATION (Weeks)	Oct 1	Oct 2	Oct 3	Oct 4	Nov 1	Nov 2	Nov 3	Nov 4	Dec 1	Dec 2	Dec 3	Dec 4	Dec 5	Jan 1	Jan 2	Jan 3	Jan 4	Feb 1	Feb 2	Feb 3	Feb 4	Feb 5	Mar 1	Mar 2	Mar 3	Mar 4	Mar 5	Apr 1	Apr 2	Apr 3	Apr 4	May 1	May 2	May 3	May 4
Initial documentation and planning																																				
Literature review	2	■	■																																	
Project proposal draft	1			■																																
Final project proposal	1				■																															
Initial architecture development																																				
Find and analyse dataset	1					■																														
Research natural language processing methods	1						■																													
Research argument extraction methods	1							■																												
Research evaluation methods	1									■																										
Design the algorithm	3										■	■	■																							
Implementation																																				
Scrape and clean dataset	3														■	■	■																			
Experiment with argument extraction methods	6																		■	■	■	■	■	■												
Evaluate model	1																								■											
Final bug fixing and testing	1																									■										
Dissertation writing																																				
Interim report	3									■	■	■																								
Dissertation	8																								■	■	■	■	■	■	■	■				
Demo preperation	4																															■	■	■	■	

7.2 Contributions and reflections

This dissertation contributes by exploring a method of argument extraction that implements social feedback. As mentioned before, the importance of user comments as valuable data cannot be ignored when it comes to the large-scale analysis of news headline data on social media platforms.

Another contribution is the creation of a news headline dataset that contains social feedback as well as a method for enhancing the original news headline.

Throughout the year, I felt that I have put in a generous amount of work since the inception of this project. I have had high hopes for this project and have hoped to create a method that would improve upon the baseline argument extraction method for news headline data. However, as the second semester started, and I began to struggle with the heavier workload. I realised that I was too naïve and underestimated the research needed to improve upon a method.
Overall, I have learned a lot from doing this project, especially about research work. I was able to learn plenty of new technologies and skills related to NLP and argument extraction as well. Although there were times where I felt like giving up, but in the end, I felt that it was a very rewarding experience overall.

8 References

[1] K. Holmqvist, J. Holsanova, M. Barthelson and D. Lundqvist, "Chapter 30 - Reading or Scanning? A Study of Newspaper and Net Paper Reading," in *The Mind's Eye*, Elsevier B.V, 2003, pp. 657-670.

[2] C. Sandianos, K. Ioannis Manousos, G. Petasis and V. Karkaletsis, "Argument Extraction from News," in *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO, Association for Computational Linguistics, 2015, pp. 56-66.

[3] M. Dusmanu, E. Cabrio and S. Villata, "Argument Mining on {T}witter: Arguments, Facts and Sources," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, 2017, pp. 2317-2322.

[4] M. Potthast, B. Stein, F. Loose and S. Becker, "Information Retrieval in the Commentsphere," in *Information Retrieval in the Commentsphere*, 2012.

[5] J. Snajder, "Social Media Argumentation Mining: The Quest for Deliberateness in," *CoRR,* vol. abs/1701.00168, 2017.

[6] C. Stab and I. Gurevych, "Identifying Argumentative Discourse Structures in Persuasive Essays," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, Doha, Qatar, Association for Computational Linguistics, 2014, pp. 46-56.

[7] A. Piotrkowicz, V. Dimitrova and K. Markert, Automatic Extraction of News Values from Headline Text, Association for Computational Linguistics, 2017.

[8] D. Hallin, "Neoliberalism, social movements and change in media systems in the late twentieth century," in *The Media and Social Theory*, Oxford, Routledge, 2008, pp. 58-73.

[9] A. Quertatani, G. Gasmi and C. Latiri, "Argued opinion extraction from festivals and cultural events on Twitter," *Procedia computer science,* vol. 126, pp. 205-213, 2018.

[10] M. Lippi and P. Torroni, "Context-Independent Claim Detection for Argument Mining," in *Proceedings of the 24th International Conference on Artificial Intelligence*, Buenos Aires, Argentina, AAAI Press, 2015, p. 185–191.

[11] Y. Zhang, G. Xu, Y. Wang, D. Lin, F. Li, C. Wu, J. Zhang and T. Huang, "A Question Answering-Based Framework for One-Step Event Argument Extraction," *IEEE Access,* vol. 8, 2020.

[12] T. H. Nguyen, K. Shirai and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications,* vol. 42, no. 24, pp. 9603-9611, 2015.

[13] X. Wang, F. Wei, X. Liu, M. Zhou and M. Zhang, "Topic Sentiment Analysis in Twitter: A Graph-Based Hashtag Sentiment Classification Approach," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management}*, Glasgow, Scotland, UK, 2011, p. 1031–1040.

[14] Z. Nurulhuda, A. Selamat and R. Ibrahim , "Hybrid sentiment classification on twitter aspect-based sentiment analysis," *Applied intelligence,* vol. 48, no. 5, pp. 1218-1232, 2017.

[15] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *European Language Resources Association (ELRA)*, Valletta, Malta, 2010.

[16] M. Henderson, P. Budzianowski, I. Casanueva, S. Coope, D. Gerz, G. Kumar, N. Mrkšić, G. Spithourakis, P.-H. Su, I. Vulić and T.-H. Wen, "A Repository of Conversational Datasets," in *Proceedings of the First Workshop on NLP for Conversational AI*, Florence, Italy, Association for Computational Linguistics, 2019, pp. 1-10.

[17] A. Akbik, D. Blythe and R. Vollgraf, "Contextual String Embeddings for Sequence Labeling," in *27th International Conference on Computational Linguistics, COLING 2018*, 2018.

[18] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research,* vol. 3, p. 993–1022, 2003.

[19] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *Transactions on Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, 1984.

[20] B. Baresch, L. Knight, D. Harp and C. Yaschur, "Friends Who Choose Your News:An analysis of content links on Facebook," *ISOJ: The Official Research Journal of International Symposium on Online Journalism,* vol. 1, pp. 65-85, 2011.

[21] T. A. Van Dijk, News as discourse, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

[22] J. Kuiken, M. Martijn, M. Spitters and M. Marx, "Effective Headlines of Newspaper Articles in a Digital Environment," *Digital Journalism,* vol. 5, no. 10, pp. 1300-1314, 2017.

[23] Y. Chen, N. Conroy and V. Rubin, "News in an online world: The need for an "automatic crap detector"," *Proceedings of the Association for Information Science and Technology,* vol. 52, no. 1, pp. 1-4, 2015.

[24] J. G. Webster, The Marketplace of Attention, Cambridge, Massachusetts: The MIT Press , 2014.

[25] M. M. U. Rony, N. Hassan and M. Yousuf, "Diving Deep into Clickbaits," *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* p. 232–239, 2017.

[26] D. Palau-Sampio, "Reference Press Metamorphosis in the Digital Context: Clickbait and Tabloid Strategies in Elpais.com," *Communication & Society,* vol. 29, no. 2, pp. 63-79, 2016.

[27] I. Habernal and I. Gurevych, "Argumentation Mining in User-Generated Web Discourse," *Computational Linguistics,* vol. 43, no. 1, pp. 125-179, 2017.

[28] H. Nguyen and D. Litman, "Extracting argument and domain words for identifying argument," in *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO, Association for Computational Linguistics, 2015, pp. 22-28.

[29] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Association for Computational Linguistics*, 2002.