Bauhaus-Universität Weimar
Faculty of Media
Degree Programme: Digital Engineering

Content and Style-Based Analysis of Persuasive Editorials

Master's Thesis

Dipendra Sharma Kafle

First Referee: Prof. Dr. Benno Stein
Second Referee: Jun.-Prof. Dr. Jan Ehlers

Submission date: 10.04.2021

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, April 10, 2021

. . . .

Dipendra Sharma Kafle

Abstract

Persuasion is abundant in the media documents; many articles (e.g., news editorials) are written to persuade the readers in the favour of a particular political ideology. Persuasion can be achieved by diverse strategies that attempt to influence the readers' minds. Most of these strategies can be categorized into content- and style-based. While the style-based strategies have been tackled in several studies, content-based ones, especially those related to argumentative topics and frames, are rarely considered. This thesis proposes methods for identifying topics and frames of argumentative articles, and use them (in addition to previously studied style features) for persuasiveness prediction. Experiments on a set of 1000 New York Times articles annotated for the persuasiveness show that style features when combined with topics and frames enhance the classifier in better prediction of persuasiveness effect.

Keywords: Natural Language Processing, Persuasion Strategies, Topic Modelling, Media Frames.

Acknowledgements

First of all, I would like to thank Prof. Dr. Benno Stein for allowing me to write thesis with the Webis Group.

I am immensely thankful to my advisers, Khalid Al-Khatib and Roxanne El-Baff, whose expertise was of great value in formulating methodology. Your guidance, insightful feedback and patient support helped me grow as a learner.

Contents

1 Introduction

Argumentation is a multi-disciplinary research field, which studies debate interactions and reasoning processes, and spans across and ties together diverse research areas such as logic and philosophy, language, psychology, and computer science(Budzynska and Reed,2019). Over the last few years, argumentation has been studied in terms of its quality assessment and analysis; several qualitative assessments across various domains such as legal documents, product reviews, scientific articles, online debates, newspaper articles, and social media have been conducted (see Chapter 3 for more details on argumentation quality and its assessment approaches).

Media has become an efficient means of influencing peoples' ideology by steering the choice of words and their presentation in the content. In the case of politics, the news media, especially the editors of the news editorials create *persuasive* arguments that favor their political stance. They state and defend a thesis that conveys their stance on a controversial topic which is usually related to the public interest(Al Khatib et al.,2016).

Persuasion is one of the main goals of argumentation, and persuasiveness has been viewed as an important dimension of argumentation quality. To produce persuasive argumentation that support their standpoints, people use different strategies such as using emotional language and dramatic appeals to beliefs and values(Nettel and Roque,2012).

This thesis studies the persuasion strategies on editorials, investigating the language used in the editorials, and particularly, the impact of selected content in addition to style on text persuasiveness. We carry out our investigation based on the following:

Task: Our tackled task in this thesis is to predict whether an editorial persuades its readers or has no effect. More specifically, an editorial is considered persuasive if it *challenges* the prior stance of readers or *reinforces* them to argue better.

The reason to focus on editorials is the shown persuasive strategies used there. Editorials are often filled with political commentary and the biggest aim of editorials is to persuade the audience in some manner (Davis,2013). Plus, the abundant presence of major style features such as emotion, sentiment, subjectivity, and various patterns of arguments make editorials a rich platform to study persuasion.

Data: We use the Webis-Editorial-Quality-18 corpus (El Baff et al.,2018) which is derived from 1000 English news editorials from New York Times newspaper. It contains the annotations regarding the persuasiveness effect. These annotations are annotated to distinguish if the editorials *challenge* the prior stance of the readers or *reinforces* their existing stance, or simply *ineffective*.

Approach: Language is a key medium for persuasion. Whether it is text or speech, persuasion is impacted by the features and quality of language. Therefore, language and its features worth to be considered in studying persuasion.

Language is not just a tool for simple communication; it has the potential of dynamic manifestations. The same event can be interpreted in various ways by using a different set of words. When one presents certain *content* with a particular *style* in an effective way, he/she is more likely to have the desired effect on the readers. The effective way (aka strategy) usually includes the structuring of ideas built upon logical argumentation, sound connection between claims and premises, and a persuasive communication style(Basave and He,2016).

Content refers to the whole body of the text. The words related to content are often nouns, regular verbs, adjectives, and adverbs, which express detailed information of the discussed topic. On the other hand, the style of the text includes word choice, tone of the sentence, sentence structure, or even type of voice used in the text. The words related to style are usually made up of pronouns, prepositions, articles, conjunctions and auxiliary verbs.(Tausczik and Pennebaker,2010).

In this thesis, we study various types of content and style features in both document and paragraph levels:

- Content features: The studied content features are (1) lemma (derived from lemmatization of the words), (2) topics, and (3) frames. We identify the topics in editorials using Latent Dirichlet Allocation (LDA) topic modeling. Furthermore, we use a BERT-based pre-trained model to develop a media frames classifier that identifies frames. For training this classifier, we use Media Frames Corpus (Card et al.,2015). The corpus includes several thousand news articles on three policy issues: same-sex marriage, tobacco, and immigration. Each article in this corpus is labeled based on 15 media frames.
- Style features: The considered style features can be grouped into mainly five style categories: (1) Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker,2010), (2) NRC Emotion&Sentiment (Mohammad and Turney,2013), (3) Webis ADUs (Al Khatib et al.,2017), (4) MPQA Arguing (Somasundaran et al.,2007) and (5) MPQA Subjectivity (Riloff and Wiebe,2003).

We develop a classifier that uses the style and content features of a given editorial as input and, as an output, it predicts its persuasiveness effect label: 'challenging', reinforcing', or 'no-effect. As a classification model, we use the Support Vector Machine (SVM).

The developed classifier is trained on the training set of Webis-Editorial-Quality-18 corpus and evaluated on the corpus test set. Based on the evaluation results (i.e., accuracy), we explore the impact of the content and style features for the prediction of persuasiveness.

Results: The results of our experiments show that the best effectiveness is achieved by combining style and content features. Among the highly important (most discriminating) features are the LIWC, NRC Emotion&Sentiment and MPQA Arguing from the style features and Lemma and Topic from the content features. The best combination was MPQA Arguing, Lemma and Topic (for liberal) and Lemma and Frame (for conservative) showing that using the topics and frames in addition to style features can improve the accuracy of the model in predicting persuasiveness.

The thesis is structured as follows:

- Chapter 2 discusses about some of the related works.
- In Chapter 3, we describe argumentation quality assessment for editorials.
- Chapter 4 deals with analysis of the content and style features for quality assessment in terms of persuasion.
- We experiment on the Webis-Editorial-Quality-18 corpus in Chapter 5 and present the best f1 scores with the best combination of features to show the importance of Topic and Frame for prediction of persuasion.
- Finally, Chapter 6 talks about the conclusion of this thesis.

2 Literature Review

In this chapter, we shed light on some of the related works. These works are based upon papers written on mainly argumentation, editorials and persuasion. Moreover, some of them have also laid the foundation of quality assessments for us in terms of style features.

In regards to persuasion and editorials, El Baff et al.2018 claims that it is evident that the news editorials affect and are said to shape public opinion. Although, being such an important source of political argumentation, the editorials do not tend to argue explicitly. Rather, they follow a subtle rhetorical strategy(El Baff et al.,2018). This theory is backed by Van Dijk1995 which found in its result that many opinions of editorials were not expressed explicitly, but implied indirectly with the help of specific factual statements filtered to empower persuasion.

Moreover, in order to study persuasion, El Baff et al.2018 considered the four dimensions defined by Virtanen and Halmari2005: (1) prior beliefs of readers, (2) prior beliefs and behaviour of authors, (3) effects of the text and (4) linguistic choices. Since these four points also act as a background for studying persuasion in this thesis, each of them are summarised briefly as followings:

(1) Prior beliefs of readers: By using the political typology quiz developed by the Pew Research Center [4], the authors considered Americans being divided into eight political groups consisting four mostly liberal and four mostly conservative ones. An additional group was also taken into consideration which included people with lesser or no interest in politics at all. However, in order to study persuasion through polarity, the paper recognised these political groups as Liberal and Conservative.

(2) Prior beliefs and behaviour of authors: The authors of this paper argued that newspapers or editors of editorials have their own set of beliefs. This can be noticed when newspapers take particular sides on controversial issues.

(3) Effects of the persuasive text: The effects of the persuasive text was monitored on basis of two major questions to the annotators: (1) If you have a different stance than the editorial, did it challenge you, making you rethink your stance? (2) If you have the same stance, did it empower you, enabling you to better defend your stance? (El Baff et al.,2018)

(4) Linguistic Choices: El Baff et al. 2020 analysed the quality in terms of how challenging or reinforcing an editorial was for the readers, given their stance.

They created a corpus with 1000 news editorials extracted from New York Times. The corpus was also annotated with annotators' political ideology and their distinguished effect. After the study of their corpus, it was found that only 1% of the editorials actually persuaded the annotators successfully. Hence, through their work, it can be concluded that annotators with different political orientation disagree on the effect significantly.

In another study, Wang et al. 2017 presented a predictive model that estimated the impact of linguistic features, the latent persuasive power of various topics and the relationship between them in the scenario of debates. In other words, they considered latent topics as content, and the linguistic features as style. Their model's combination of content and style predicted audience-adjudicated winners with a significant 74% accuracy, whereas linguistic features alone could only achieve an accuracy of 66%. Thus, we also use topics as one of the content features for our experiments.

Moreover, El Baff et al. 2020 also presented different types of style features and their importance in terms of persuasion. On the basis of editorials found on aforementioned NY Times corpus, they compared style and content based classifiers in collation to ideology and corresponding effects. In addition to finding about conservative readers being resistant to liberal NY Times style and style having bigger impact on the liberal readers as compared to content, the authors, most importantly, found that the content and style based classifier performed better than style-based or content-based classifiers alone. Since this paper includes major style features from various studies and deals with persuasion, we use style features it contains by considering the similarity in our tasks and goal.

The results of aforementioned papers, favouring combination of content and style features over a single one for better prediction, are also supported by Basave and He 2016. They studied the effect of a speaker's argumentation style in influencing an audience in supporting their candidature. They modelled the influence index of each candidate which was based on their corresponding standings in the polls released prior to the debate. They created a classifier that ranked speakers in terms of their relative influence by using a combination of content and persuasive argumentation features. Although this paper is based on debates instead of editorials, the result still confirmed persuasive argumentation style affected such indices and played an important role in predicting a speaker's influence rank while combined to content.

In the past, some pivotal research has been done in the pursuit of defining style features. One of such works introduces Linguistic Inquiry and Word Count (LIWC). Tausczik and Pennebaker 2010 defined LIWC as a transparent text analysis program which counts words in psychologically meaningful categories such as thinking style, emotionality, individual differences, social relationships and attentional focus. These categories were followed by other sub-categories such as positive and negaive emotions, status dominance and social heirarchy, honesty and deception, social cordination and group processes.A more refined version of LIWC was later brought in by Pennebaker et al.

2015, which introduced 15 distinguished dimensions of LIWC. The same version was also adapted by El Baff et al. 2020 as one of the style features.

Following are the 15 dimensions of LIWC (Pennebaker et al.,2015) along with their brief description: (1) Language Metrics : This contains words per sentence, long words that have more than 6 letters and dictionary words. (2) Function Words: This includes personal pronouns, impersonal pronouns, articles, prepositions, auxilary verbs, common adverbs, conjunctions and negations. (3) Other Grammar: It consists of regular verbs, adjectives, comparatives, interrogatives, numbers and quantifiers. (4) Affect Words: This is composed of positive and negative emotion followed by anxiety, anger and sadness. (5) Social Words: This refers to family and friends. (6) Cognitive Processes: It comprises of insight, cause, discrepancies, tentativeness, certainty and differentiation. (7) Perceptual Process: It includes seeing, hearing and feeling. (8) Biological Processes: This relates to body related terms regarding health/illness, sexuality and ingesting. (9) Drives and Needs: This one refers to affiliation, achievement, power, reward focus and risk focus. (10) Time Orientations: It deals with focus related to past, present and future. (11) Relativity: It deals with motion, space and time. (12) Personal Concerns: Work, leisure, home, money religion and death are its constituents. (13) Informal Speech: It includes all the swear words, netspeak, assent, non-fluencies and fillers. (14) All Punctuation: General punctuation marks such as periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses and other punctuation comes under this one. (15) Summary Variables: This dimension has four of its own sub-dimensions.

The four variables representing (15) are: (a) Analytical Thinking: This refers to the degree of formality used in the language. For instance; formal and logical language scores high degree whereas, narrative language scores low (Pennebaker et al.,2014). (b) Clout: It measures leadership, relative social status and confidence in text(Kacewicz et al.,2014). (c) Authenticity: It deals with the degree to which a person reveal himself/herself authentically(Newman et al.,2003). (d) Emotional tone: This final variable scales the polarity in the tone. Score more than 50 refers to positive emotional tone, and lesser than 50 means negative emotional tone(Cohn et al.,2004).

In addition to LIWC, El Baff et al. 2020 also implemented NRC EmotionSentiment(Mohammad and Turney,2013) style feature. NRC includes a set of English words mapped with emotions such as anger, fear and disgust. The words are also mapped with negative and positive sentiment polarities. NRC feature is measured in terms of count of words corresponding to each association.

More on style features, Al Khatib et al. 2017 identified evidence types in 300 news editorials obtained from The Guardian, Fox News and Al Jazeera. They categorised the evidence into 3 types: (1) Statistical: In this type, text mentions the facts as result of studies, data analyses or other sorts of quantitative research. (2) Testimonial: In this case, text states or quotes an argument, a concept or a proposal presented by some expert, authority, witness or similar kinds of honourable and trustworthy entities, (3) Anecdotal: The text talks about personal experience of the author, a concrete example, an instance, a specific event, or similar(Al Khatib et al.,Al Khatib et al.). El Baff et al. 2020 referred to this feature as Webis ADUs, which were represented as the count of words corresponding to each evidence type.

In another study, Somasundaran et al. 2007 studied relation between dialog structure and expression of the opinion in a scenario of multi-party discourse in meetings. As a result, they created a lexicon that contains different patterns of argument like authority, assessments, emphasis and doubts. El Baff et al. 2020 also applied this lexicon feature in which each lexicon was represented by the count of the respective feature pattern in an editorial. The feature was coined as MPQA Arguing.

Including all of the aforementioned style features, the final one used in this thesis is based on number of subjective and objective sentences in editorials. In 2003, Riloff and Wiebe 2003 constructed a bootstrapping classifier which learned linguistically rich extraction patterns for subjective expressions. Subjective expressions refers to segments of the language which indicate opinions or subjectivity. These patterns were used to identify subjective sentences in the text. The classifier was also applied later in El Baff et al. El Baff et al. to count the number of subjective and objective sentences in editorials. They referred to this feature as MPQA Subjectivity.

In the context of content feature, papers in the past have approached it in their own unique way. For instance; as mentioned above, Wang et al. 2017 considered topics as the content feature. Whereas, El Baff et al. 2020 used ’lemma 1- to 3- grams’ as their content feature. Lemma can be understood as a canonical form, a dictionary form or a citation form of a set of words(Zgusta,2012). Papers such as Wang et al. 2017 and El Baff et al. 2020 have already showed the significance of using content in combination with style. Therefore, we too adapt content features to study more on this sort of combination and ultimately to use them for better prediction of persuasion.

Apart from topics and lemmas, an interesting feature to study in terms of content is media frames. Framing is a persuasion strategy in which media manipulates information on controversial policy issues by emphasising on certain favourable parts of facts while cutting out other aspects which can be damaging to their arguments. One of the initial and major research work on media frames is (Boydstun et al.,2014) where they introduced 15 dimensions of media framing. Later (Card et al.,2015)created Media Frames Corpus based on those 15 dimensions which included several thousand news articles on mainly three policy issues: same-sex marriage, tobacco and immigration. More information on these are provided in chapter 4.2.

Besides content and style features, Lukin et al. 2017 presented a hypothesis that certain personality types may be more or less convinced by particular styles of argument. Similarly, they added that some personalities may be effected by emotional persuasiveness, whereas others may be inclined to factual arguments. They further went on to report that persuasion and convincing potential were affected by personality factors. For instance, conscientious, open and agreeable people were more convinced by emotional arguments.

The study of personality traits has been broadly based upon the the "Big Five" personality traits which was initially presented by Goldberg 1990. The same can also be found in the paper El Baff et al. 2018 where they discovered correlations among the effect of editorials, political ideologies and personality traits of readers.

Finally, by gathering valuable concepts, proven hypotheses and compelling conclusions from all of the related works mentioned above, we move forward to further research and experiment on them in pursuit of uncovering pivotal relations with persuasion. Moreover, by combining topics and frames as new content features with the style features in El Baff et al. 2020, we further aim to improve the accuracy score of their classifier that predicts the persuasive effect of editorials.

3 Argumentation Quality

Argumentation quality can be understood as the audience's subjective perception of the arguments in the persuasive message as strong and cogent on the one hand versus weak and specious on the other(Petty et al.,1981). In order to describe argumentation quality, some of its prime aspects are essential to be considered. For instance; the factors that make an argument rational or unacceptable, persuasive or far-fetched, and so on.

We focus on the type of argumentation that consists of linguistic choices such as certain content combined with certain style of presentation or expression. Quality assessment is a way to analyse such argumentation. In this chapter, we discuss about argument quality assessment in general and also in context of editorials which deals with persuasive effectiveness.

Apart from effectiveness, there are various other dimensions to argumentation quality. Section 3.1 describes the overview of argumentation quality assessment. Whereas, in section 3.2 we discuss especially about the assessment based on editorials.

3.1 Argumentation Quality Assessment Overview

In order to assess argumentation quality, different approaches has been presented in the past. These approaches have their own views which can be summarised into two categories:(1) Theory-based views of quality assessment and (2) Practical-based views of quality assessment (Wachsmuth et al.,2017).

As per theory, a logical and convincing argument has acceptable premises which correspond to the conclusion and is sufficient to draw the conclusion(Johnson and Blair,2006); nonetheless in practice, researchers find it difficult to assess such quality dimensions for real-life arguments(Habernal and Gurevych,2016).

While theoretical-based approach encapsulates quality dimensions such as cogency, effectiveness and reasonableness (Wachsmuth et al.,2017), practical-based approach considers aspects such as quality assessments in relative terms(Cabrio and Villata,2012), ranking arguments in terms of persuasion(Wachsmuth et al.,2017) and relevance, and so on. Table 1 and Table 2 represent the theory-based and practical-based quality dimensions presented by (Wachsmuth et al.,2017) below.

As we can see on Table 1 and Table 2, the three main dimensions: Logic, Rhetoric and Dialectic summarises argumentation quality. Logic refers to the rationality of arguments, often being based on facts. Rhetoric belongs to indirect persuasive effect of arguments. And finally, Dialectic deals with reasonableness of argumentation. (Wachsmuth et al.,2017). In this thesis, we deal with editorials where we particularly focus on Ef-

Aspect	Quality Dimensions	Summary of Dimensions
Logic	**Cogency**	Argument has acceptable,relevant, and sufficient premises.
Dialectic	Local acceptability	Premises worthy of being believed.
Logic	Local sufficiency	Premises enough to draw conclusion.
Logic	Local relevance	Premises support/attack conclusion.
Rhetoric	**Effectiveness**	Argument persuades audience.
Rhetoric	Credibility	Makes author worthy of credence.
Rhetoric	Emotional appeal	Makes audience open to arguments.
Rhetoric	Clarity	Avoids deviation from the issue, uses unambiguous language.
Rhetoric	Appropriateness	Language proportional to the issue, supports credibility.
Rhetoric	Arrangement	Argues in the right order.
Dialectic	**Reasonableness**	Argument is (globally) acceptable, relevant, and sufficient.
Dialectic	Global acceptability	Audience accepts use of argument.
Dialectic	Global relevance	Argument helps arrive at agreement.
Dialectic	Global sufficiency	Enough rebuttal of counterarguments.
	Overall Quality	Argumentation quality in total.

Table 1: The 15 theory-based quality dimensions (grouped by bold-lettered high dimensions) by (Wachsmuth et al.,2017)

fectiveness and Persuasiveness belonging to the Rhetoric dimension. However, in our case, we consider Effectiveness in Table 1 and Persuasiveness in Table 2 in terms of linguistic choices with respect to argumentation quality. Moreover, we approach these with a dialectical perspective, as described in the upcoming section 3.2.

Aspect	Quality Dimensions
Logic	Evidence.
Logic	Level of support
Logic	Sufficiency
Rhetoric	Argument strength
Rhetoric	Evaluability
Rhetoric	Global Ccoherence
Rhetoric	Organization
Rhetoric	Persuasiveness
Rhetoric	Prompt adherence
Rhetoric	Thesis clarity
Dialectic	Acceptability
Dialectic	Convincingness
Dialectic	Prominence
Dialectic	Relevance

Table 2: Practical assessment of quality dimensions presented by (Wachsmuth et al.,2017)

3.2 Argumentation Quality Assessment for Editorials

An editorial is an article in the newspaper which presents the opinion of newspaper's editor on various issues related to politics, business, healthcare, economy and so on. The prime objective of the editorial is to change the belief, orientation or stance of readers whose prior orientation is different than that of the editorial's editors. If the reader happens to be of the same orientation, an editorial aims at reinforcing the reader at arguing better on the given issue.

We analyse linguistic choices with respect to argumentation quality in editorials based on a corpus 'Webis-Editorial-Quality-18' by El Baff et al. 2018. They defined the dialectical perspective of the cited paper and use their corpus to further study on quality assessment.

The quality of an editorial is mainly based upon two dimensions: (1) the prior beliefs of the reader and (2) the effect of the text (El Baff et al.,2018). The prior belief, in this case, refers to the political ideology of the readers. We have already discussed two major political ideologies: Liberal and Conservative in Chapter 2, In this unit, we talk about the constituents of these ideologies and how they can be determining argumentation quality for editorials.

Liberal ideologies consist of solid liberals, opportunity democrats, disaffected democrats, and devout and diverse. Whereas, Conservative ideologies include core conservatives, country first conservatives, market sceptic republicans, and new era enterprisers. Since it is difficult to comprehend and access the general prior beliefs of all the readers, political ideology serves as a proxy to model the reader's prior beliefs(El Baff et al.,2018).

Next, the effect of the news editorial can be measured in terms of how challenging or how reinforcing the editorial is(El Baff et al.,2018). An editorial is labelled as challenging if it makes readers reevaluate their prior stance on the given issue. This does not mean the readers will change their stance. In the end, the readers may or may not discard their prior stance, belief or orientation.In contrast, if the editorial empowers the prior stance of the readers on the given issue, it is labelled as reinforcing.

Challenging and Reinforcing can be further sub-divided into five categories in order to encapsulate the magnitude of the effect(El Baff et al.,2018). The five categories are mentioned as follows:

- Strongly challenging: The editorial firmly makes the readers reevaluate their prior stance and rethink why their belief is correct.
- Somewhat challenging: The readers find at least some new and noteworthy information opposite to their stance in the editorial.

- No effect: The readers does not find any new or noteworthy information that supports or defies their prior stance.
- Somewhat challenging: The editorial contains at least some new and noteworthy information that supports or empowers the readers' prior stance.
- Strongly reinforcing: The readers are strongly empowered by the editorial with better arguments in support of their stance.

For all the experiments in this thesis, we base our work on the aforementioned corpus and the stated analogy of persuasive effects.

4 Content and Style Features for Quality Assessment

As we have already described about Content and Style features in Chapter 1 and 2, we directly proceed to implement them in this chapter. We start with analysis of topics, followed by frames.

4.1 Topic Modelling

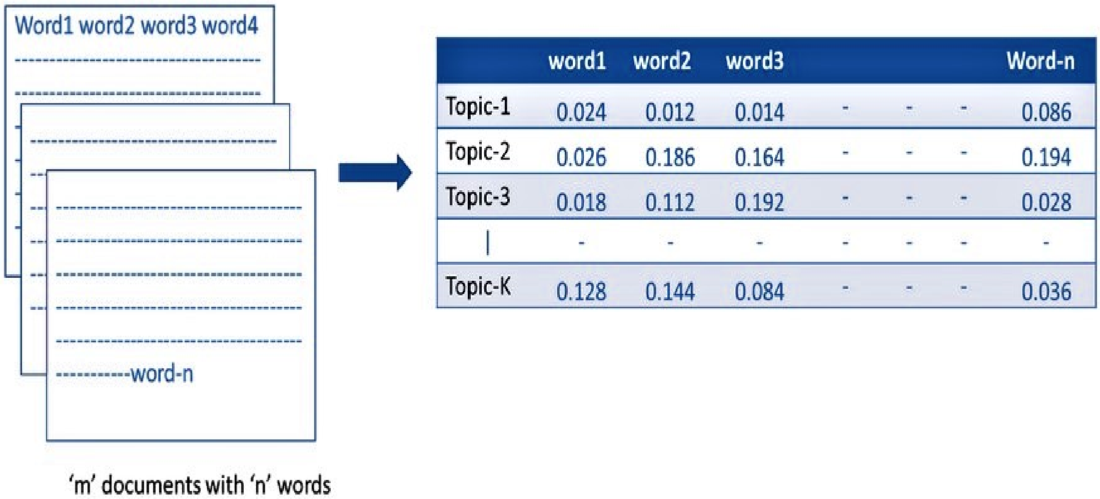
4.1.1 Introduction to Topic Modelling

In natural language processing, a topic model refers to kind of statistical model that discovers the abstract "topics" which occur in a collection of documents or texts. Topic modelling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body (Wikipedia).

Topic Modelling uses model based on mathematical framework that examines a set of texts (referred as documents), compares statistics of the words in each document and finally discovers the potential topics out of it. There are different algorithms for getting topics, such as Latent Dirichlet Analysis (LDA), Latent Semantic Analysis (LSI) and Non-negative Matrix Factorisation (NMF) and so on. Out of these, we choose LDA to conduct our topic modelling.

LDA is a cutting edge technique for content analysis that is designed to automatically organise large archives of documents based on latent topics, measured as patterns of word (co-)occurrence. It is a useful tool for analysing trends and patterns in news content in large digital news archives relatively quickly(Jacobi et al.,2016).

The main idea behind LDA is: each document can be described by a distribution of topics and each topic can be described by a distribution of words[3].In other words, it identifies topics within the documents and map documents to those topics. As the name suggests, 'Latent' refers to the finding of hidden topics from the document. Similarly, 'Dirichlet' represents LDA's model assumption that the distribution of topics in a document and the distribution of words in topics are both Dirichlet distributions, where Dirichlet distribution creates n positive numbers (a set of random vectors X1. . . Xn) that add up to 1. Lastly, 'Allocation' refers to the distribution of topics in the document.



	word1	word2	word3				Word-n
Topic-1	0.024	0.012	0.014	-	-	-	0.086
Topic-2	0.026	0.186	0.164	-	-	-	0.194
Topic-3	0.018	0.112	0.192	-	-	-	0.028
|	-	-	-	-	-	-	-
Topic-K	0.128	0.144	0.084	-	-	-	0.036

Figure 1: The probability estimates for topic assignment to words[1]

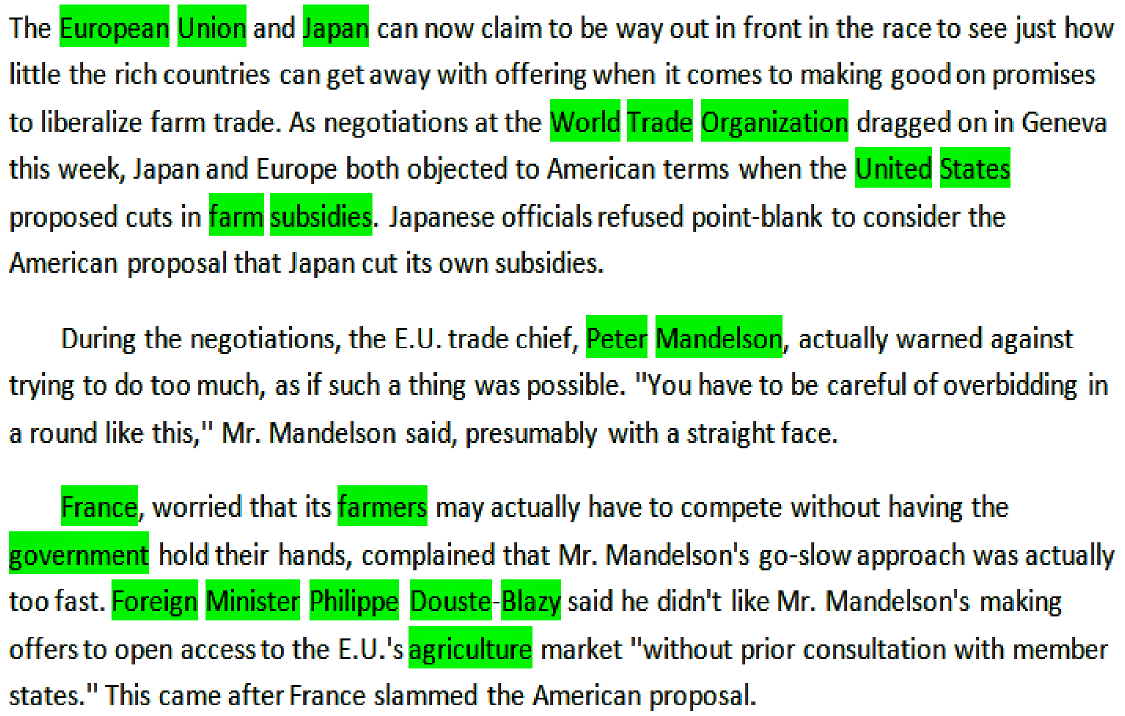
Furthermore, the assignment of each word in the document to different topics is in terms of conditional probability estimates. As displayed in Figure 1, the value in each cell indicates the probability of a word w*i* belonging to topic t*j*. Here, 'i' and 'j' are the word and topic indices respectively. While, LDA ignores the order of occurrence of words and the syntactic information, it treats documents just as a collection of words or a bag of words.

LDA is based on Bayes estimation which also takes the priors for the parameters into account. If those priors are accurate, it includes a lower risk in parameter estimation for generating meaningful topics. In contrast to others, LDA also excels when we have documents with just few words because it depends on the prior to obtain a more reasonable guess about the topics. Since, we perform topic modelling also at sentence level and our corpus is news archives, we prefer LDA. Nonetheless, we also try experimenting another popular topic modelling technique called Non-Negative Matrix Factorization (NMF), just to ensure we select the right one for our task.

Non-Negative Matrix Factorization is an unsupervised technique. NMF factorises high-dimensional vectors into a lower-dimensional ones. These lower-dimensional vectors and their coefficients both are non-negative. By using the original matrix (A), it provides two matrices (W and H). While, W is the topics that it found, H is the coefficients (weights) for those topics. It adjusts the initial values of W and H so that the product approaches A until either the approximation error converges or the max iterations are reached.

Experiment:

Based on the Webis corpus, we did manual annotations for a sample list of documents and followed a qualitative approach at understanding which algorithm works better at prediction of topics. For the annotation part, we highlighted the words in the editorials that possess the potential to be the topic/topics of the editorial. In other words, we manually annotated 50 editorials from 999 editorials (as two editorials out of 1000 had the same content), and stored the words which describe the title or theme of the editorials. An instance of this is shown in Figure 2.



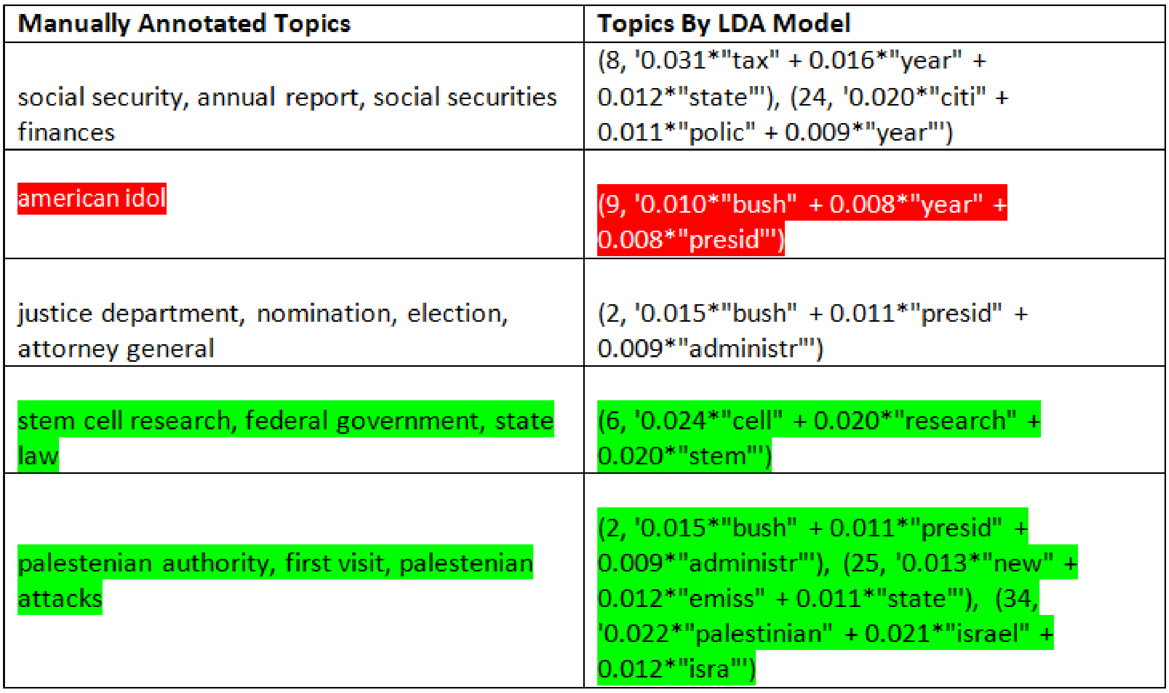
The European Union and Japan can now claim to be way out in front in the race to see just how little the rich countries can get away with offering when it comes to making good on promises to liberalize farm trade. As negotiations at the World Trade Organization dragged on in Geneva this week, Japan and Europe both objected to American terms when the United States proposed cuts in farm subsidies. Japanese officials refused point-blank to consider the American proposal that Japan cut its own subsidies.

During the negotiations, the E.U. trade chief, Peter Mandelson, actually warned against trying to do too much, as if such a thing was possible. "You have to be careful of overbidding in a round like this," Mr. Mandelson said, presumably with a straight face.

France, worried that its farmers may actually have to compete without having the government hold their hands, complained that Mr. Mandelson's go-slow approach was actually too fast. Foreign Minister Philippe Douste-Blazy said he didn't like Mr. Mandelson's making offers to open access to the E.U.'s agriculture market "without prior consultation with member states." This came after France slammed the American proposal.

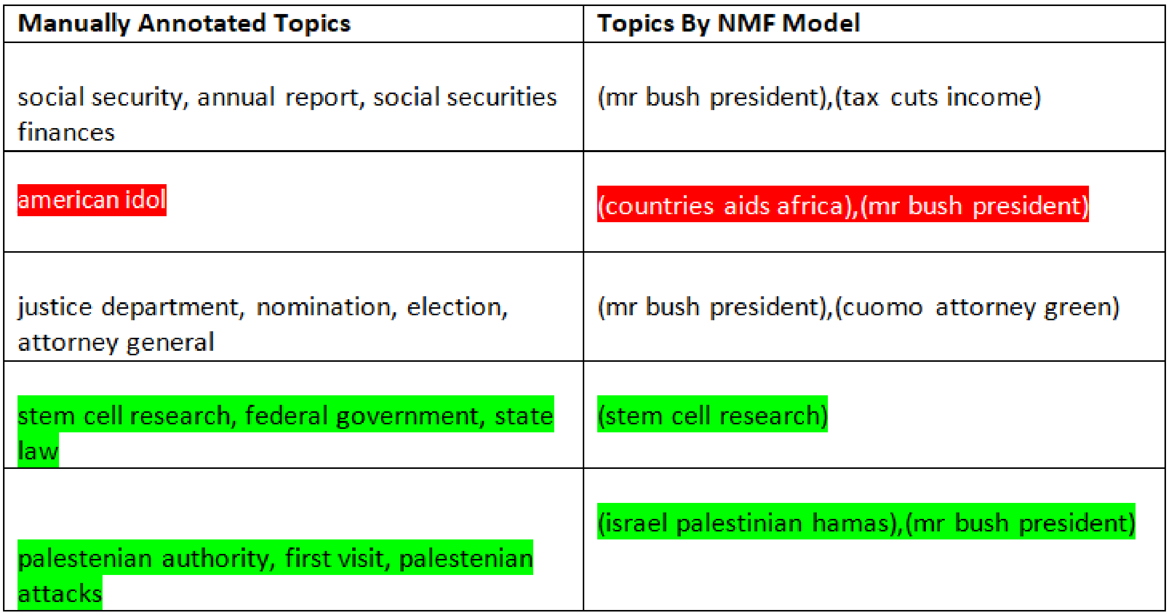
Figure 2: Topics annotated manually for an editorial. The highlighted words in green represent potential topics.

In the next step, we performed LDA topic modelling and generated 50 topics. We related each of the editorials to their corresponding dominant topic. Then, we compared those LDA topics with our manual annotation ones. Figure 3 shows a side-by-side comparison of the topics obtained through the two approaches.



Manually Annotated Topics	Topics By LDA Model
social security, annual report, social securities finances	(8, '0.031*"tax" + 0.016*"year" + 0.012*"state"'), (24, '0.020*"citi" + 0.011*"polic" + 0.009*"year"')
american idol	(9, '0.010*"bush" + 0.008*"year" + 0.008*"presid"')
justice department, nomination, election, attorney general	(2, '0.015*"bush" + 0.011*"presid" + 0.009*"administr"')
stem cell research, federal government, state law	(6, '0.024*"cell" + 0.020*"research" + 0.020*"stem"')
palestenian authority, first visit, palestenian attacks	(2, '0.015*"bush" + 0.011*"presid" + 0.009*"administr"'), (25, '0.013*"new" + 0.012*"emiss" + 0.011*"state"'), (34, '0.022*"palestinian" + 0.021*"israel" + 0.012*"isra"')

Figure 3: This figure represents comparison between manual and LDA topics. Green indicates 'matched', red indicates 'no match' and white indicates 'somewhat matched' between LDA topics and manual ones



Manually Annotated Topics	Topics By NMF Model
social security, annual report, social securities finances	(mr bush president),(tax cuts income)
american idol	(countries aids africa),(mr bush president)
justice department, nomination, election, attorney general	(mr bush president),(cuomo attorney green)
stem cell research, federal government, state law	(stem cell research)
palestenian authority, first visit, palestenian attacks	(israel palestinian hamas),(mr bush president)

Figure 4: This figure represents comparison between manual and NMF topics. Green indicates 'matched', red indicates 'no match' and white indicates 'somewhat matched' between NMF topics and manual ones

Similarly, we carried out same steps for NMF as well (illustrated in Figure 4). As a result, out of 50 editorials (chosen at random), we got matching topics on 39 editorials with LDA and 31 editorials with NMF. In other words, LDA performed better than

NMF when compared to our manually annotated topics. Thus, this result too suggested that LDA seemed a better fit for our study.

4.1.2 Analysis of Topics and Persuasion Effect

In order to analyse the effect of topics on persuasion, we study the relation between each topic and the editorials associated with it. Moreover, we correspond those 999 editorials to 6000 annotations with three effect labels: challenging , no-effect and reinforcing.

Experiment:

Before linking all the aspects such as topics, editorials and effect labels together, it was crucial to pay some attention to the generation of topics via LDA model. We used LDAModel from Gensim (Řehřek et al.,2011) library. Since, number of topics is expected as an input for this model, we decided to check coherence scores for different number of topics.

Coherence, or in our case, topic coherence measures the degree of semantic similarity between high scoring words in the topics. On the basis of this measurement, it provides scores to each of the topics. These scores helps to distinguish between topics whose associated words are semantically better linked to each other. We used CoherenceModel from Gensim library to achieve this score. Figure 5 shows the number of topics vs. coherence score where bigger coherence score represent potential of generating better topics. On checking coherence for a range of 20-50 topics, we found highest score when the number of topics was 27. Therefore, we decided to use 27 topics for further experiments.

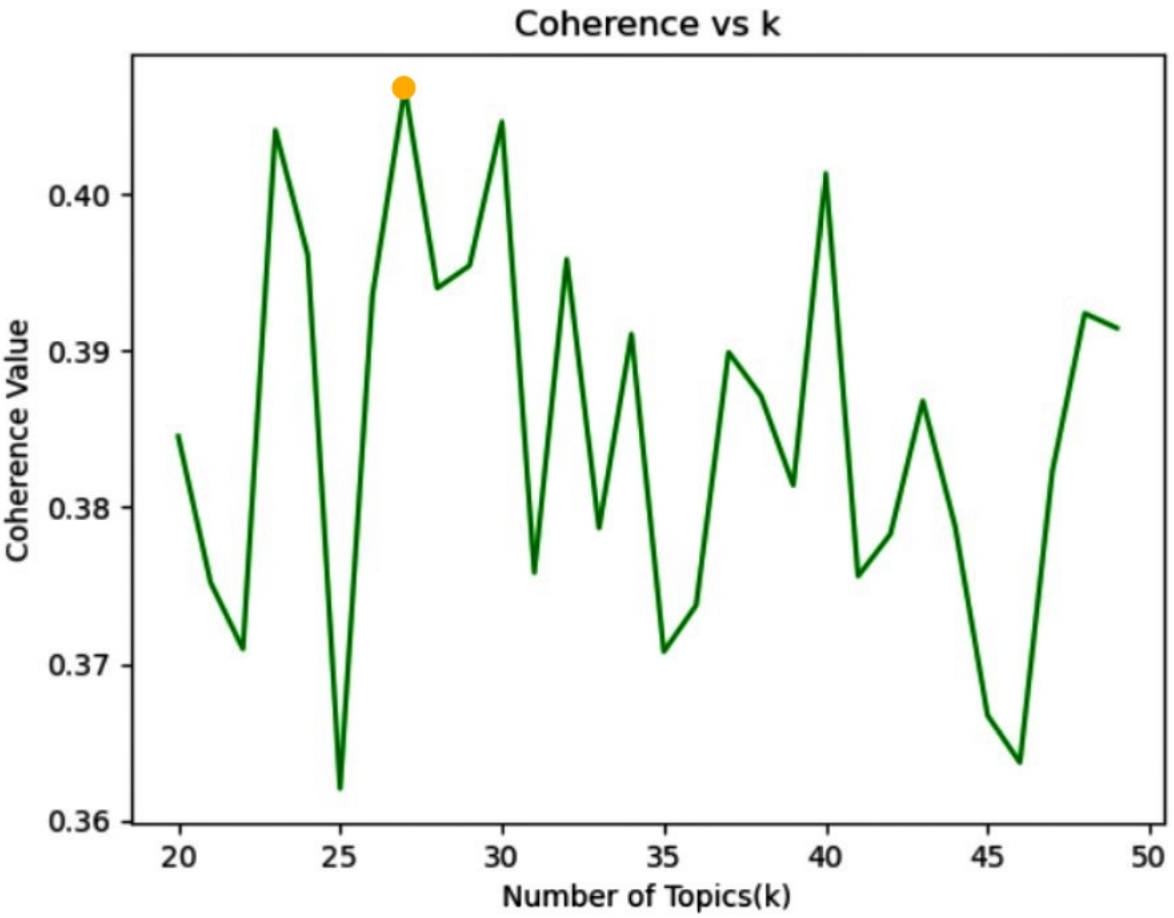


Figure 5: From a range of 20-50 topics, the coherence model provides highest coherence value (score) when the number of topics used is 27 (as shown by the yellow dot).

4.1.3 Topic Modelling at Article Level

For the next step, we studied the distribution of editorials associated to each topic at article. As mentioned earlier, LDA model generates numerous topics for a document. We found the dominant topics for each document based on the weights that each of the topics carries. Finally, we linked the editorials to their corresponding dominant topics. The resulting distribution is shown in Figure 6. As the figure suggests, topics such as immigration, President Bush and Iraq, New reforms in state were frequently used by the New York Times newspaper; whereas, topics such as budget of the state, divorce regarding women and senate policies on drugs were less frequent.

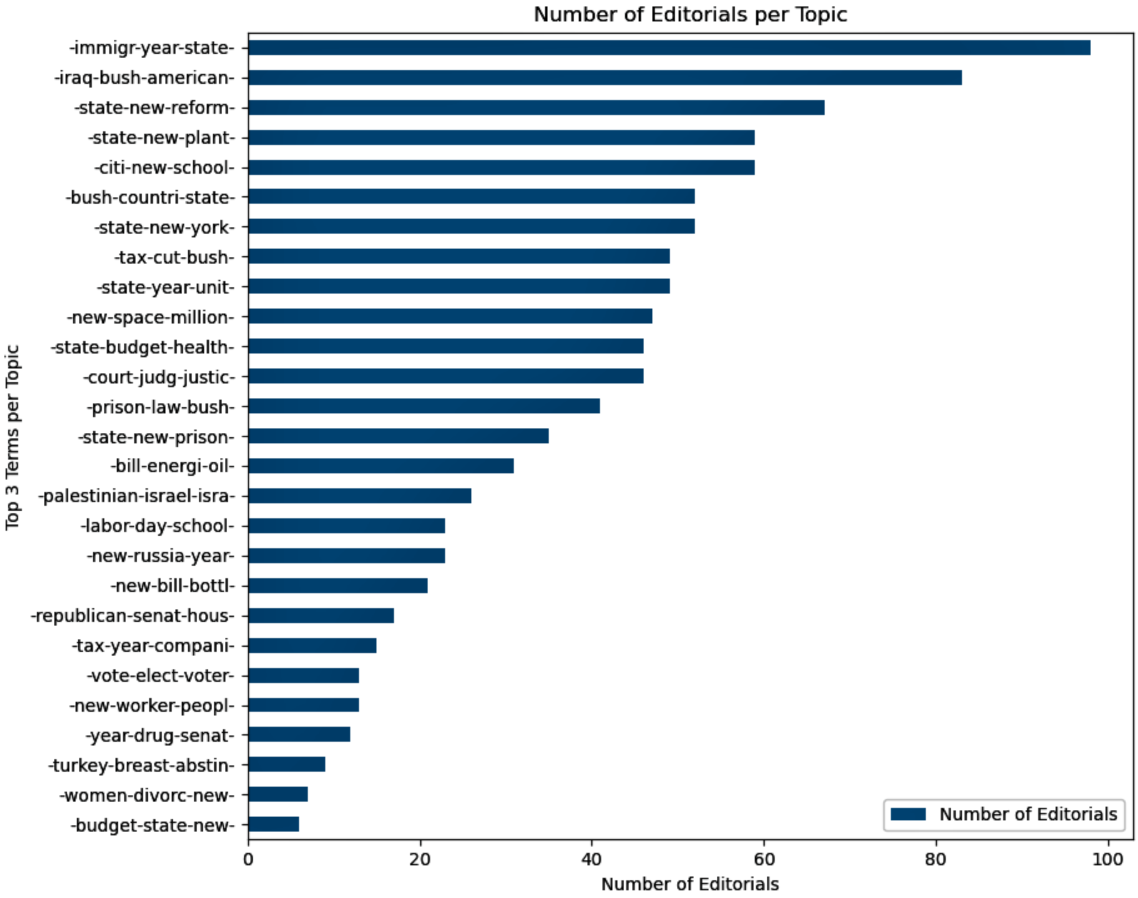


Figure 6: The distribution of 999 editorials across 27 topics generated by LDA model. Topics are represented on Y-axis by their corresponding top 3 terms.

After having some knowledge on the distribution of editorials across 27 topics, we visualised the relation between topics and effect labels in order to study persuasion. Figure 7 describes how each topic is linked to each of the effect labels.

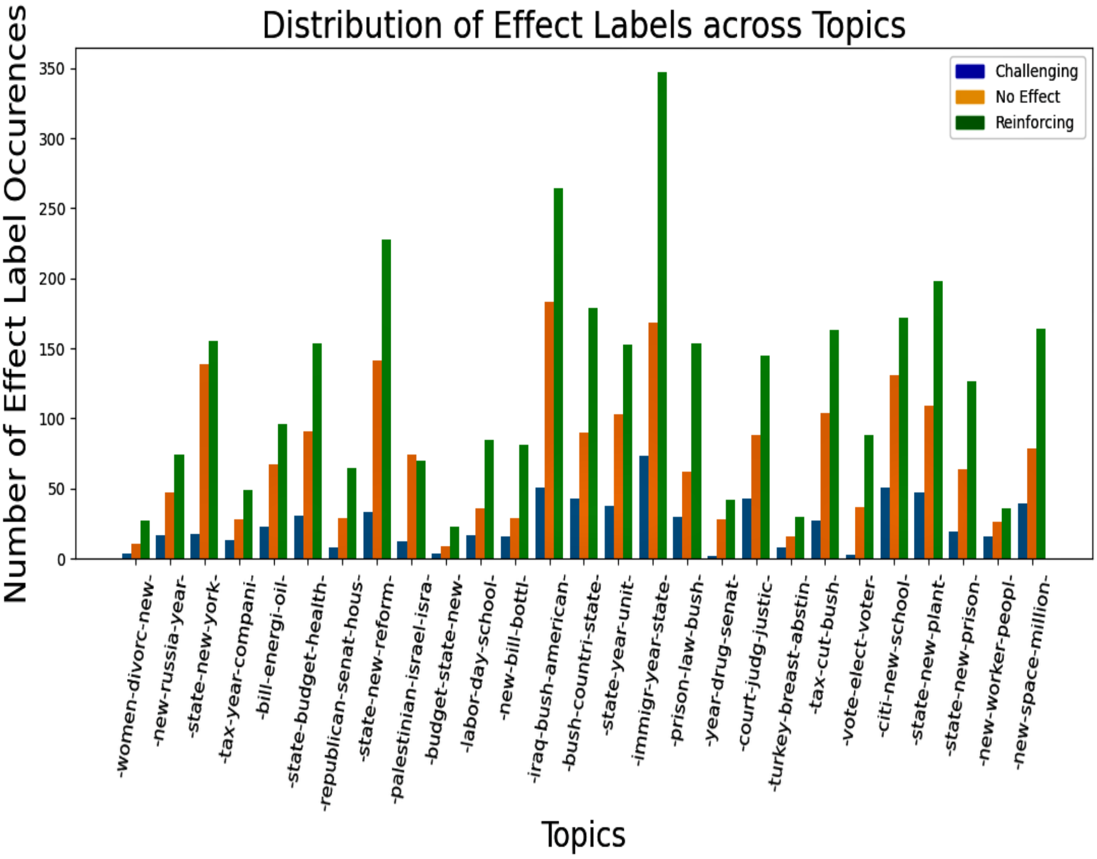


Figure 7: The distribution of effect labels across 6000 annotations represented by 27 topics. Topics are represented on X-axis by their corresponding top 3 terms.

From Figure 7, we already know that editorials with the effect label ’Reinforcing’ are larger in number as compared to other labels. Considering this fact, we can justify the distribution of effect labels in Figure 7 where we found lesser number of editorials for ’Challenging’ and ’No Effect’ as compared to ’Reinforcing’.

Moreover, with the help of Figure 7, we discovered which topics challenged or reinforced annotators the most. Some of the major findings are listed below:

- Immigration, Iraq and Bush, and Education were the top 3 topics that challenged the annotators.
- Immigration, Iraq and Bush, and State Reforms were top 3 topics that reinforced the annotators.
- Immigration, Iraq and Bush, and New York were top 3 topics which had no effect on annotators.
- Drugs, Election and Divorce were the least challenging topics.
- State Budget, Turkey and Divorce were the least reinforcing topics.

Out of 6000 annotations conducted on 999 editorials, it can be observed that larger number of editorials are associated to topics like Bush and Iraq, Immigration and State Reform.So, it is obvious that larger part of the annotations fall under these topics. However, keeping this fact in mind, we analysed the differences between the bars(representing each topic) in Figure 7. Although being top topics for 'No Effect'; Immigration, Iraq and Bush both challenged and reinforced annotators in most of the annotations as compared to other topics. Therefore, we argue these topics played a major role in persuasion and argumentation quality. On the other hand, topics such as Divorce and Labour were less frequently used for persuasion among articles by New York Times newspapers.

Thus, based on the obtained information from the corpus, we conclude that topics falling under the category of Immigration, Foreign Policy and Politics are often frequently and effectively used for persuasion. Whereas, topics related to Economy, Jurisdiction and Labour are not the prime 'go-to' topics for creating greater persuasion among the readers.

4.1.4 Topic Modelling at Paragraph Level

After studying topics in article level, we moved to paragraph level. We took the same corpus for this task as well where we divided editorials into their constituent paragraphs and considered them as documents. The main goal of this task was to find out whether we get more accurate topics with words that can better summarise the documents.

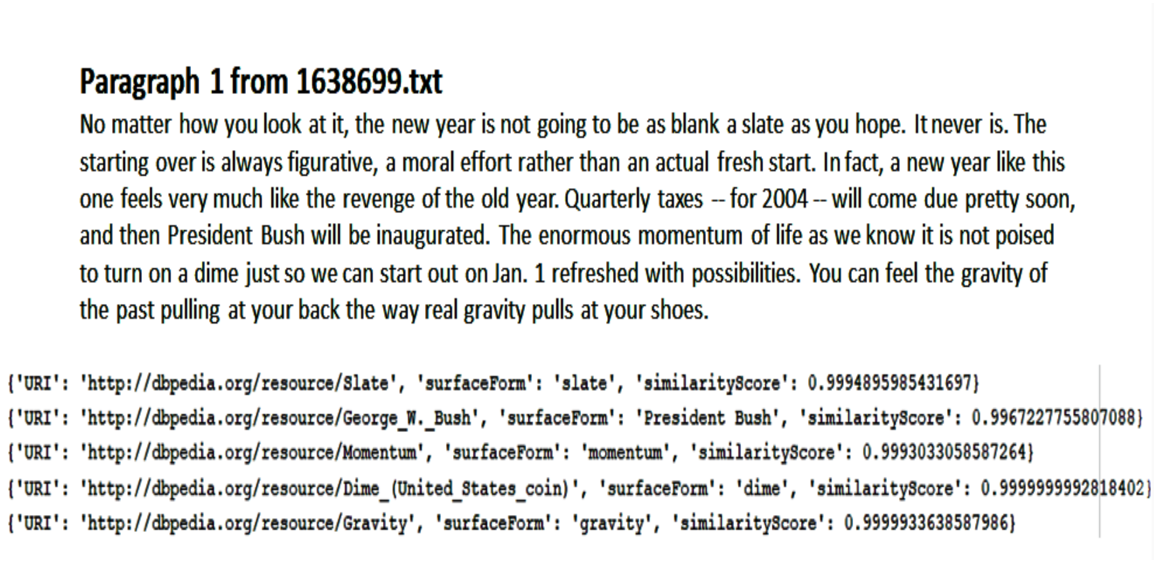
Along with LDA topic model, we also used DBpedia Spotlight which is a tool that automatically annotates mentions of DBpedia resources in text. It also provides a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia ([2]). The annotation basically provides 3 helpful information:

1 **URL**: Url for the correspoding DBpedia resource.

2 **SurfaceForm**: The annotated word chosen and linked by Dbpedia.

3 **similarityScore**: A score of similarity between URL and SurfaceForm.

In other words, Dbpedia annotates the document, compares and links the words that are mentions of DBpedia resources in the document. As a result, we get some keyowrds which describe the concept of the document. In our case, it provides us with the topics of the paragraphs. It was used in order to compare and evaluate the performance of LDA model at paragraph level. Figure 8 and Table 3 illustrate one of the instances of how we performed the experiment and compared the results.



Paragraph 1 from 1638699.txt

No matter how you look at it, the new year is not going to be as blank a slate as you hope. It never is. The starting over is always figurative, a moral effort rather than an actual fresh start. In fact, a new year like this one feels very much like the revenge of the old year. Quarterly taxes -- for 2004 -- will come due pretty soon, and then President Bush will be inaugurated. The enormous momentum of life as we know it is not poised to turn on a dime just so we can start out on Jan. 1 refreshed with possibilities. You can feel the gravity of the past pulling at your back the way real gravity pulls at your shoes.

```
{'URI': 'http://dbpedia.org/resource/Slate', 'surfaceForm': 'slate', 'similarityScore': 0.9994895985431697}
{'URI': 'http://dbpedia.org/resource/George_W._Bush', 'surfaceForm': 'President Bush', 'similarityScore': 0.9967227755807088}
{'URI': 'http://dbpedia.org/resource/Momentum', 'surfaceForm': 'momentum', 'similarityScore': 0.9993033058587264}
{'URI': 'http://dbpedia.org/resource/Dime_(United_States_coin)', 'surfaceForm': 'dime', 'similarityScore': 0.9999999992818402}
{'URI': 'http://dbpedia.org/resource/Gravity', 'surfaceForm': 'gravity', 'similarityScore': 0.9999933638587986}
```

Figure 8: Illustration of Dbpedia annotations where first paragraph from editorial 1638699.txt is annotated. Dbpedia links the text to five of its resouces along with surfaceform and similarity score.

As shown in Figure 8, Dbpedia produces 5 topics from 5 resources for the given paragraph. We further ran our pre-trained LDA model on the same paragraph and received 7 topics with 3 topic describing terms each. The results of both the approaches are shown side by side in Table 4.

LDA Topics	Dbpedia Topics
tax, cut, pay	dime
bush, **presid**, hous	gravity
state, new, senate	momentum
immigr, can, bill	**President Bush**
american, iraq, militari	slate
year, percent, billion	
hous, administration, **bush**	

Table 3: Comparison between topics generated by LDA and Dbpedia. The topics are arranged according to the weights (LDA) and similarity score (Dbpedia) in descending order. Bold lettered topics are the common topics between two approaches.

The paragraph in Figure 8 talks about new year and how it is not a fresh start. This is followed by including other factors such as taxes and inauguration of president Bush. While LDA model made 'tax' as the dominant topic as per the highest weight, Dbpedia considered 'dime' as the best describing word for the paragraph. This is certainly not accurate. But, if we consider other topics with lower weights as well, LDA includes topics such as tax, bush, state, senate, immigration, iraq, military and so on. It is clear that only a couple of topics describe the paragraph. However, the description is not adequately accurate as it misses 'new year' and misdirects us with unrelated terms such as immigration, iraq and so on. On the other hand, Dbpedia better identifies the keywords in the texts. But, most of these keywords are unable to help us infer what the paragraph is about in actual.

4.1.5 Comparing The Results of Topic Modelling at Each Level

In order to compare the results at each level, we followed the same steps as described in Figure 8 and Table 3. By choosing several paragraphs from different editorials at random, we manually analysed the topics generated at each granular level. As a result, we found that LDA topic modelling at paragraph level did not produce as good results as at article level. At paragraph level, only few topic keywords were matched with that of Dbpedia. Although, the performance at article level was better, it does not mean that LDA model performed poorly at paragraph level. Therefore, later in Chapter 5, we also use topics of paragraphs alongside article level topics to check if it helps classifier predict the persuasion effect better.

4.2 Frames Modelling

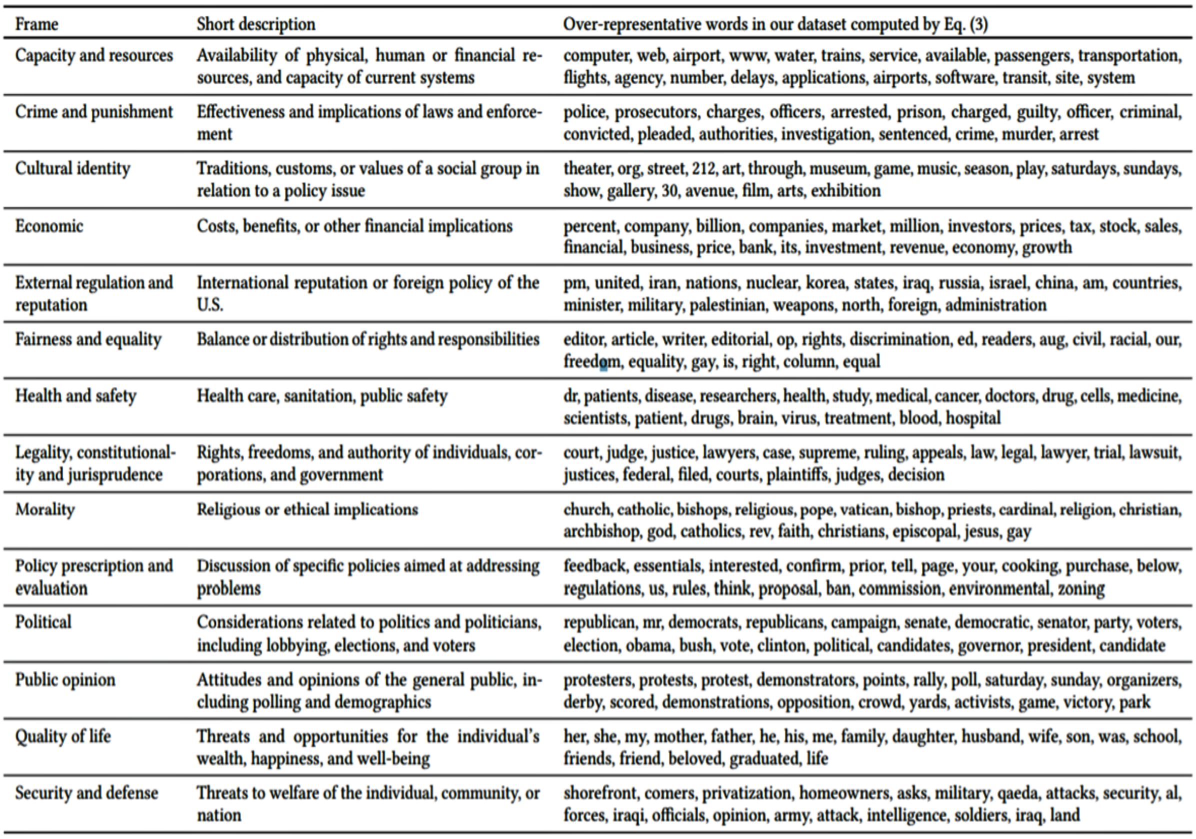
We have already briefly discussed about Frames in Chapter 1 and 2. In this chapter, we discuss in detail what frames are, how we model or generate them and how we can use them to study persuasion and even predict persuasion.

4.2.1 Introduction to Frames

As, we already know, framing is a persuasion strategy in which media manipulates information on controversial policy issues by emphasising on certain favourable parts of facts while cutting out other aspects which can be damaging to their arguments(Boydstun et al.,2014). The different dimensions or categories of such framing are known as frames.

We get 15 frames from Media Frames Corpus (Card et al.,2015) which summarises the all the dimensions of frames presented by Boydstun et al. 2014. These 15 dimensions include several thousand news articles on mainly three policy issues: same-sex marriage, tobacco and immigration. Out of 15, one of the dimensions is labelled as ’other’ which is used when the text does not belong to any of the 14 dimensions. So, practically Media Frames Corpus (MFC) consists of 14 distinct types of frames representing various issues in media.

Figure 9 displays all the dimensions of frames in MFC with their corresponding short description.



Frame	Short description	Over-representative words in our dataset computed by Eq. (3)
Capacity and resources	Availability of physical, human or financial resources, and capacity of current systems	computer, web, airport, www, water, trains, service, available, passengers, transportation, flights, agency, number, delays, applications, airports, software, transit, site, system
Crime and punishment	Effectiveness and implications of laws and enforcement	police, prosecutors, charges, officers, arrested, prison, charged, guilty, officer, criminal, convicted, pleaded, authorities, investigation, sentenced, crime, murder, arrest
Cultural identity	Traditions, customs, or values of a social group in relation to a policy issue	theater, org, street, 212, art, through, museum, game, music, season, play, saturdays, sundays, show, gallery, 30, avenue, film, arts, exhibition
Economic	Costs, benefits, or other financial implications	percent, company, billion, companies, market, million, investors, prices, tax, stock, sales, financial, business, price, bank, its, investment, revenue, economy, growth
External regulation and reputation	International reputation or foreign policy of the U.S.	pm, united, iran, nations, nuclear, korea, states, iraq, russia, israel, china, am, countries, minister, military, palestinian, weapons, north, foreign, administration
Fairness and equality	Balance or distribution of rights and responsibilities	editor, article, writer, editorial, op, rights, discrimination, ed, readers, aug, civil, racial, our, freedom, equality, gay, is, right, column, equal
Health and safety	Health care, sanitation, public safety	dr, patients, disease, researchers, health, study, medical, cancer, doctors, drug, cells, medicine, scientists, patient, drugs, brain, virus, treatment, blood, hospital
Legality, constitutionality and jurisprudence	Rights, freedoms, and authority of individuals, corporations, and government	court, judge, justice, lawyers, case, supreme, ruling, appeals, law, legal, lawyer, trial, lawsuit, justices, federal, filed, courts, plaintiffs, judges, decision
Morality	Religious or ethical implications	church, catholic, bishops, religious, pope, vatican, bishop, priests, cardinal, religion, christian, archbishop, god, catholics, rev, faith, christians, episcopal, jesus, gay
Policy prescription and evaluation	Discussion of specific policies aimed at addressing problems	feedback, essentials, interested, confirm, prior, tell, page, your, cooking, purchase, below, regulations, us, rules, think, proposal, ban, commission, environmental, zoning
Political	Considerations related to politics and politicians, including lobbying, elections, and voters	republican, mr, democrats, republicans, campaign, senate, democratic, senator, party, voters, election, obama, bush, vote, clinton, political, candidates, governor, president, candidate
Public opinion	Attitudes and opinions of the general public, including polling and demographics	protesters, protests, protest, demonstrators, points, rally, poll, saturday, sunday, organizers, derby, scored, demonstrations, opposition, crowd, yards, activists, game, victory, park
Quality of life	Threats and opportunities for the individual's wealth, happiness, and well-being	her, she, my, mother, father, he, his, me, family, daughter, husband, wife, son, was, school, friends, friend, beloved, graduated, life
Security and defense	Threats to welfare of the individual, community, or nation	shorefront, comers, privatization, homeowners, asks, military, qaeda, attacks, security, al, forces, iraqi, officials, opinion, army, attack, intelligence, soldiers, iraq, land

Figure 9: Description of 14 frames in Media Frames Corpus (Card et al.,2015) with their short description and words that are over-representative of those frames.

4.2.2 Development of Frames Classifier

One of the studies that deals with frames is Kwak et al.2020 which studied frames and built a classifier that predicts frames for the text. Unfortunately, we could not use their classifier due to lack of access to the source code repository. Therefore, we built our own classifier with the help of labelled dataset in Media Frames Corpus.Figure 10 shows the distribution of frames in the corpus.

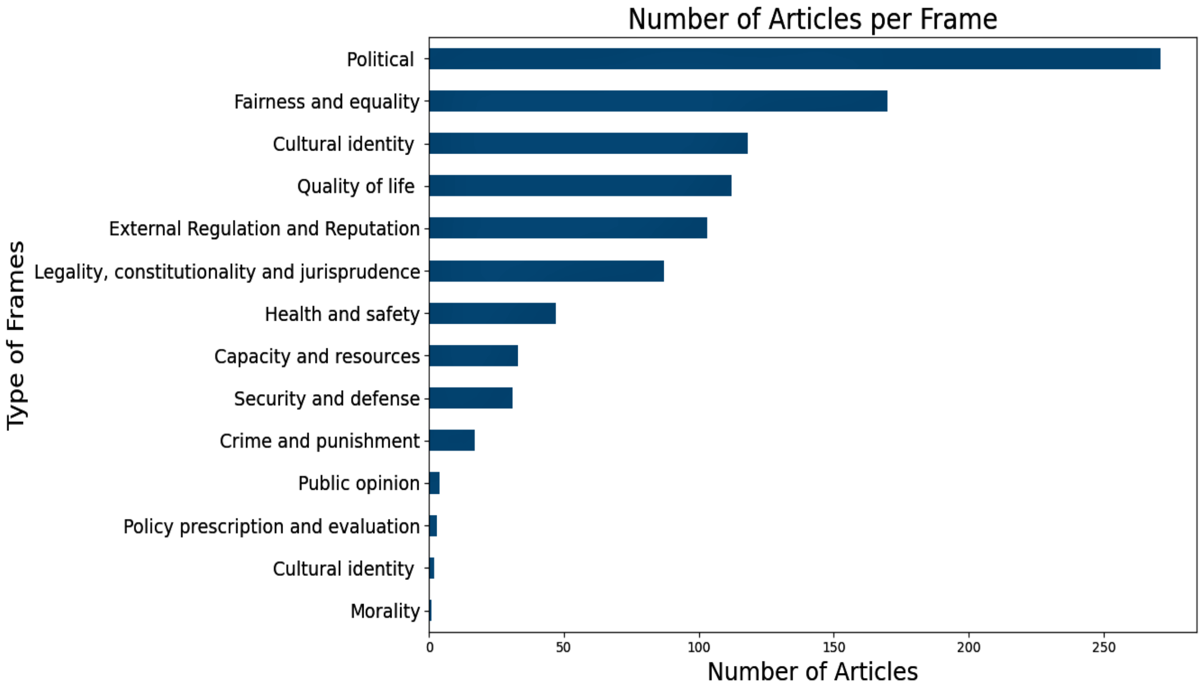


Figure 10: Distribution of 14 frames accross 11,014 articles in Media Frames Corpus.

Our classifier is based on BERT-base model that helps us classify documents into different frames. This model is a pre-trained model which means that it was already trained on a bigger dataset. So, our job was to train the model on our relatively smaller MFC dataset so as to make it fit for our task. In other words, we fine-tuned the Bert-base model.

In order to fine-tune the model, we obtained the labelled MFC corpus with its 3 different sub-categories: Same-sex marriage, Immigration and Tobacco. Altogether, we got 11,014 articles that we trained and tested using 10-fold Cross Validation. We also used following hyper-parameters for BERT-base classifier:

- Batch size: It defines the number of samples to iterate through before updating the internal model parameters. A batch is like a for-loop that iterates over one or more samples and makes predictions. At the end of each batch, the predictions are compared to the expected output variables and an error is estimated. We used a batch size of 32 for our model.

- Training epoch: This refers to the number of times the learning algorithm iterates through the entire training dataset. We set this to 3.

- Maximum sequence number: The maximum sequence length of Bert is 512, which means only 512 tokens from the text will be converted into token ids and fed into the classifiers as tensors. By considering the computation time and memory, we set this to 128.

- Learning rate: This hyper-parameter determines the magnitude of change in the model with respect to the estimated error each time the model weights are updated. We used a learning rate of 2e-5.

As a result, we got average macro-f1 score of 0.47 from 10-fold cross validation. In the next section, we further analyse the corpus and study the correctly and incorrectly predicted frames as a part of error analysis.

4.2.3 Error Analysis

In order to conduct error analysis, we randomly selected two samples of 30 articles each from correctly predicted and incorrectly predicted articles. Figure 11 shows us the distribution of incorrectly predicted articles. Moreover, in Figure 12, we show the pattern of keywords that play a major role in the wrong predictions by the classifier. We did the same for correctly predicted ones as well which is shown in Figure 13 and Figure 14.

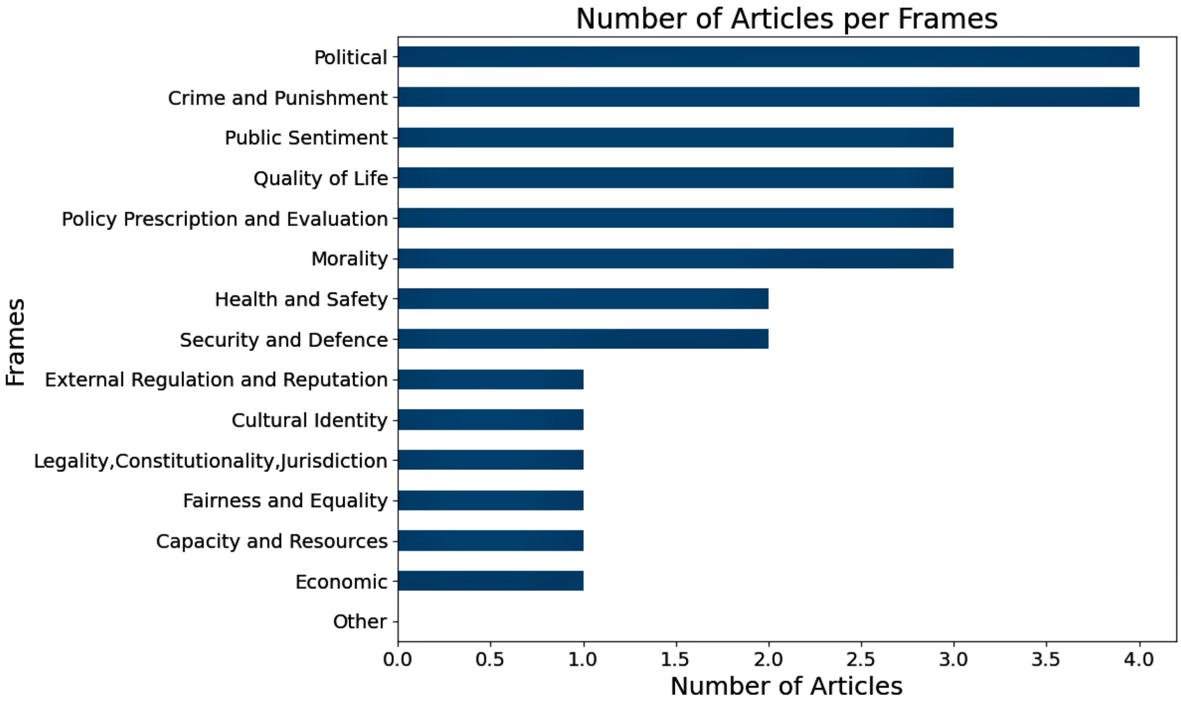
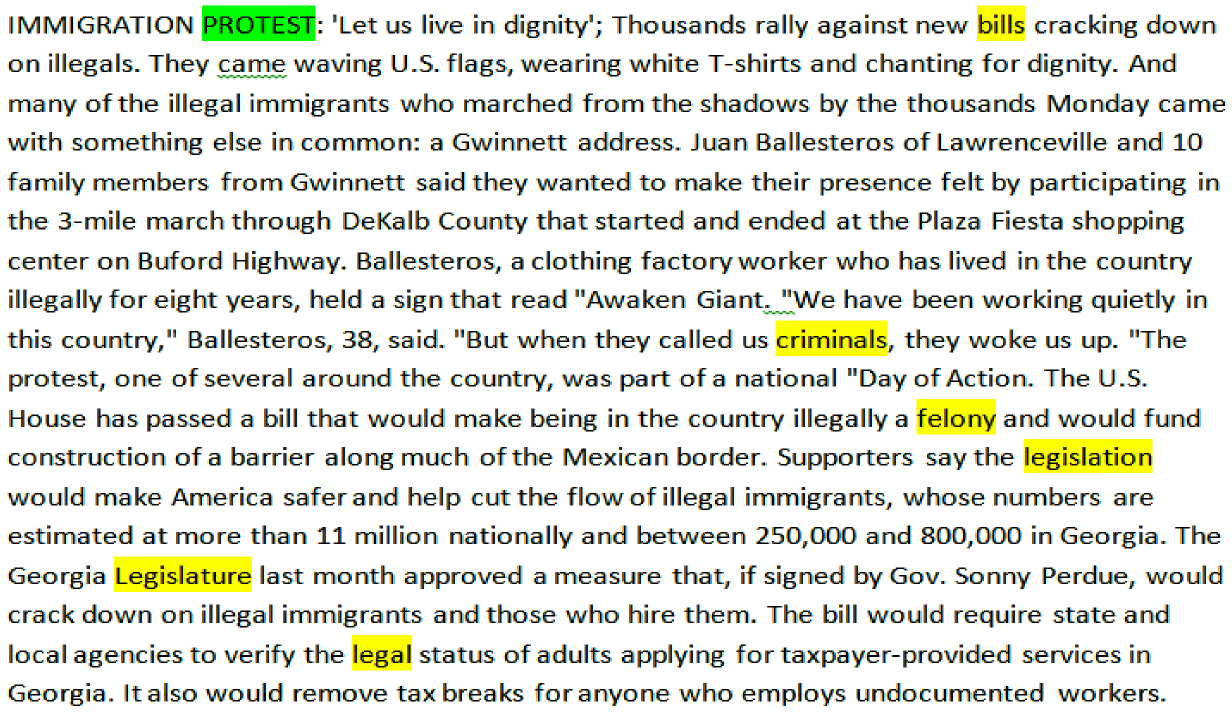


Figure 11: Distribution of 15 frames across 30 incorrectly predicted articles in Media Frames Corpus.



IMMIGRATION PROTEST: 'Let us live in dignity'; Thousands rally against new bills cracking down on illegals. They came waving U.S. flags, wearing white T-shirts and chanting for dignity. And many of the illegal immigrants who marched from the shadows by the thousands Monday came with something else in common: a Gwinnett address. Juan Ballesteros of Lawrenceville and 10 family members from Gwinnett said they wanted to make their presence felt by participating in the 3-mile march through DeKalb County that started and ended at the Plaza Fiesta shopping center on Buford Highway. Ballesteros, a clothing factory worker who has lived in the country illegally for eight years, held a sign that read "Awaken Giant. "We have been working quietly in this country," Ballesteros, 38, said. "But when they called us criminals, they woke us up. "The protest, one of several around the country, was part of a national "Day of Action. The U.S. House has passed a bill that would make being in the country illegally a felony and would fund construction of a barrier along much of the Mexican border. Supporters say the legislation would make America safer and help cut the flow of illegal immigrants, whose numbers are estimated at more than 11 million nationally and between 250,000 and 800,000 in Georgia. The Georgia Legislature last month approved a measure that, if signed by Gov. Sonny Perdue, would crack down on illegal immigrants and those who hire them. The bill would require state and local agencies to verify the legal status of adults applying for taxpayer-provided services in Georgia. It also would remove tax breaks for anyone who employs undocumented workers.

Figure 12: The article is about public sentiment. Word highlighted in green is the ideal keyword that the classifier should pick and predict 'public sentiment'. But instead, it picks words highlighted in yellow and gives wrong prediction as Legality, Constitutionality and Jurisdiction.

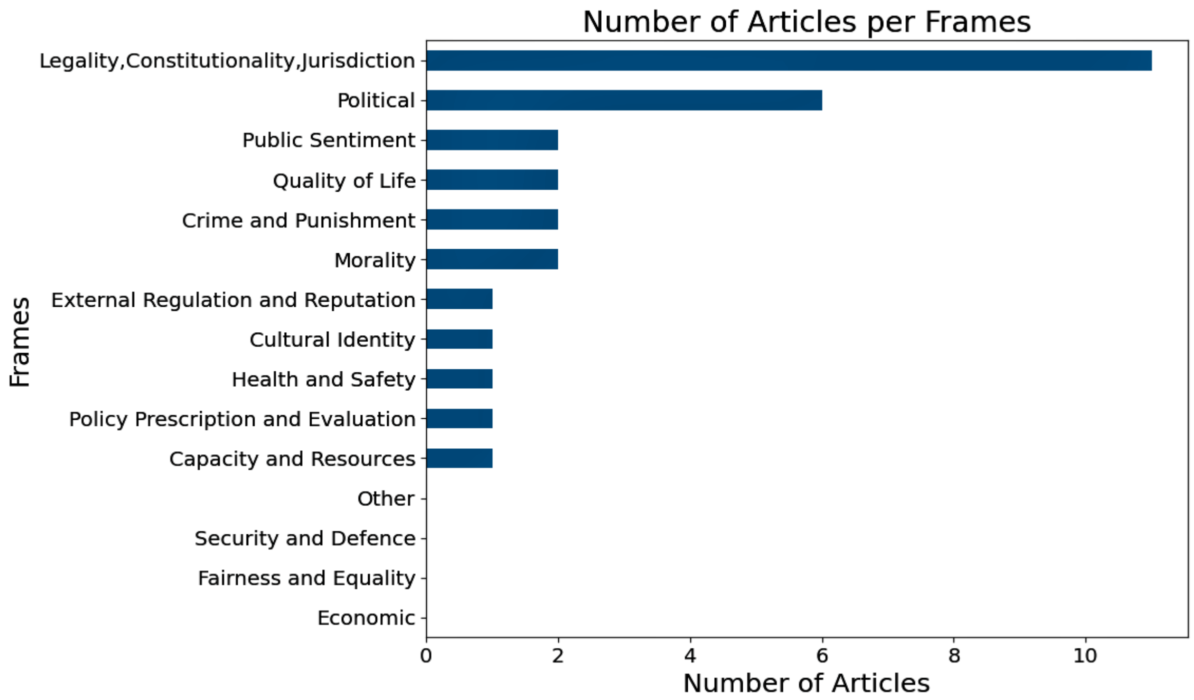
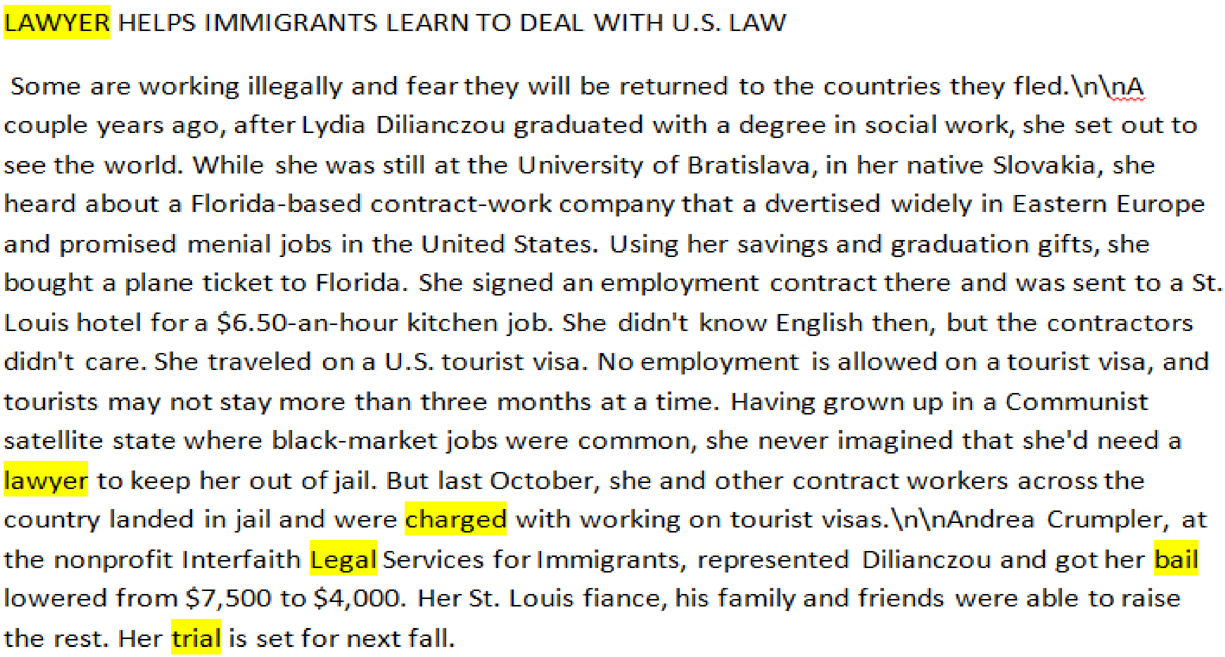


Figure 13: Distribution of 15 frames accross 30 correctly predicted articles in Media Frames Corpus.



LAWYER HELPS IMMIGRANTS LEARN TO DEAL WITH U.S. LAW

Some are working illegally and fear they will be returned to the countries they fled.\n\nA couple years ago, after Lydia Dilianczou graduated with a degree in social work, she set out to see the world. While she was still at the University of Bratislava, in her native Slovakia, she heard about a Florida-based contract-work company that a dvertised widely in Eastern Europe and promised menial jobs in the United States. Using her savings and graduation gifts, she bought a plane ticket to Florida. She signed an employment contract there and was sent to a St. Louis hotel for a $6.50-an-hour kitchen job. She didn't know English then, but the contractors didn't care. She traveled on a U.S. tourist visa. No employment is allowed on a tourist visa, and tourists may not stay more than three months at a time. Having grown up in a Communist satellite state where black-market jobs were common, she never imagined that she'd need a lawyer to keep her out of jail. But last October, she and other contract workers across the country landed in jail and were charged with working on tourist visas.\n\nAndrea Crumpler, at the nonprofit Interfaith Legal Services for Immigrants, represented Dilianczou and got her bail lowered from $7,500 to $4,000. Her St. Louis fiance, his family and friends were able to raise the rest. Her trial is set for next fall.

Figure 14: The article is about Legality, Constitutionality,Jurisdiction. Words highlighted in yellow are correctly picked by the classifier and predict the target frame correctly.

From the aforementioned process and figures, we try to tackle the question: What leads to correct and incorrect predictions. Our findings are listed below:

- The texts with explicit (frame) keywords were easier to detect than the ones which lack a good number of explicit keywords.

  Example

  Political : This frame had more number of explicit keywords such as senate, party, president, election, vote and so on.

  This frame contained lesser number of explicit keywords like protest and rally.

  Example

- The frames with lesser number of articles had lesser number of tokens (keywords) to train.

  Frames with lesser articles: external regulation and reputation, capacity and resources,fairness and equality and other.

  Frames with more articles: political, economic, legality, constitutionality, jurisdiction, public prescription and evaluation.

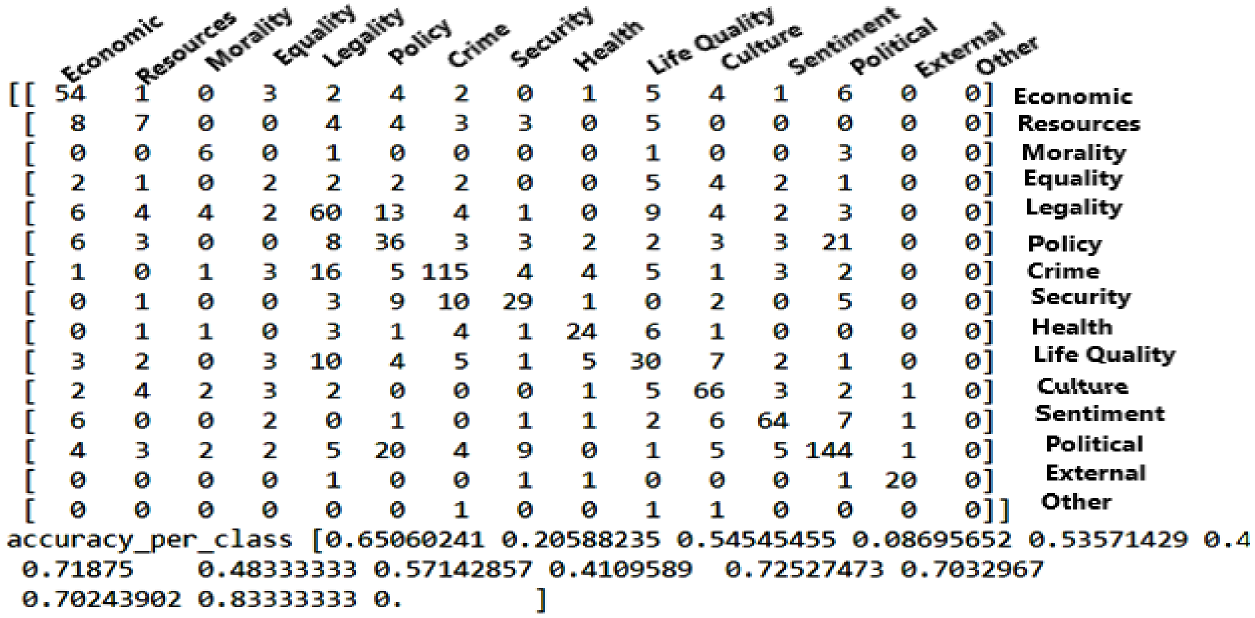
- Some frames were a bit more abstract than the others which made it difficult for the classifier to predict accurately.

  Example

Frames like public opinion, quality of life, morality and cultural identity had abstract meanings with very few number of keywords in the text. On the other hand, we found plenty of keywords related to other frames but not related to the actual dominant/primary frame of the article.

4.2.4 Classifying Frames at Article Level

In this part of the experiment, we optimised the existing classifier to improve the accuracy of prediction. We used the same dataset and used the default hyper-parameters values, except the learning rate. We changed learning rate from 2e-5 to 5e-5. Then, we tested the model using of 10-fold cross validation. As a result, we got an accuracy of micro-f1 = 0.60 and macro-f1 = 0.48, which is a slight improvement as compared to the accuracy of previous model (mentioned in section 4.2.2). Figure 15 shows the improved prediction of the classifier in a confusion matrix.



```
[[ 54   1   0   3   2   4   2   0   1   5   4   1   6   0   0]
 [  8   7   0   0   4   4   3   3   0   5   0   0   0   0   0]
 [  0   0   6   0   1   0   0   0   0   1   0   0   3   0   0]
 [  2   1   0   2   2   2   2   0   0   5   4   2   1   0   0]
 [  6   4   4   2  60  13   4   1   0   9   4   2   3   0   0]
 [  6   3   0   0   8  36   3   3   2   2   3   3  21   0   0]
 [  1   0   1   3  16   5 115   4   4   5   1   3   2   0   0]
 [  0   1   0   0   3   9  10  29   1   0   2   0   5   0   0]
 [  0   1   1   0   3   1   4   1  24   6   1   0   0   0   0]
 [  3   2   0   3  10   4   5   1   5  30   7   2   1   0   0]
 [  2   4   2   3   2   0   0   0   1   5  66   3   2   1   0]
 [  6   0   0   2   0   1   0   1   1   2   6  64   7   1   0]
 [  4   3   2   2   5  20   4   9   0   1   5   5 144   1   0]
 [  0   0   0   0   1   0   0   1   1   0   0   0   1  20   0]
 [  0   0   0   0   0   0   1   0   0   1   1   0   0   0   0]]
accuracy_per_class [0.65060241 0.20588235 0.54545455 0.08695652 0.53571429 0.4
 0.71875    0.48333333 0.57142857 0.4109589  0.72527473 0.7032967
 0.70243902 0.83333333 0.        ]
```

Figure 15: Confusion matrix showing the distribution of correctly and incorrectly predicted class labels (frames). The abbreviated names of the frames with their corresponding index is shown in the right.

As we can see, most of the classes have bigger numbers situated along the diagonal and most of the classes have accuracy above 0.50. This suggests that model can be used for further experiments. However, just to be sure, we again selected a sample of articles at random from the internet (The New York Times) and compared the manually annotated primary frame of the articles with the predicted frames provided by the classifier. In this analysis, out of 30 articles, we found 19 articles whose primary frame matched our manual annotated ones. The ones that did not match belonged to 'abstract frames' which is already described in section 4.2.

4.2.5 Classifying Frames at Span Level

Besides articles, Media Frames Corpus also includes annotations for frames at span level. A span is a part of the article that was marked by annotators for being a decisive part in the text which helped them choose and label the corresponding frames. A span can be a sentence, a paragraph or somewhere in between sentences and paragraphs. We obtained 22,377 spans from 11,014 articles in the corpus. Moreover, we trained our model on spans with the same hyper-parameters.

We split the dataset into Train-Validation-Test set assigning 60%, 20% and 20% respectively. As a result, we got an accuracy of micro-f1: 0.41 and macro-f1: 0.39. The classifier performed poorly at spans level. As an alternative, we decided to train the classifier at sentence level as well.

4.2.6 Classifying Frames at Sentence Level

We converted the spans into sentences in the following manner:

1 If a span was a sentence, we took it in our new sentence-level dataset and labelled the frame, same as the frame of the corresponding span.

2 If a span ranged between more than one sentences, we selected all those sentences for our new dataset and labelled them with the same corresponding frame of the span.

3 If a sentence had more than one span, we selected the frames from all of the corresponding spans.

The conversion of spans into sentences gave us a new dataset with 135,212 labelled sentences. Once again, we trained our classifier on sentence-level dataset. The dataset was split into train, validation and test set consisting of 60%, 20% and 20% dataset respectively. We used following hyper-parameters for the model:

- Batch size: 32
- Training epoch: 3
- Maximum sequence number: Since a sentence has lesser words than an article. We set this hyper-parameter to 25 in contrast to 128 used previously.
- Learning rate: 5e-5

We obtained the accuracy of micro-f1: 0.41 and macro-f1: 0.37. As these scores were not good enough to finalize the model for future experiments, we decided to tweak 2 hyper-parameters in order to try achieving higher accuracy score. We set 'Maximum sequence number' back to 128 because it would contain all the words/tokens of even the longer sentences. Lastly, we tried several number of 'Learning rate' ranging from 2e-5 to 6e-5, and selected the one that caused the lowest training loss. Figure 16 shows different learning rates compared to their corresponding training loss.

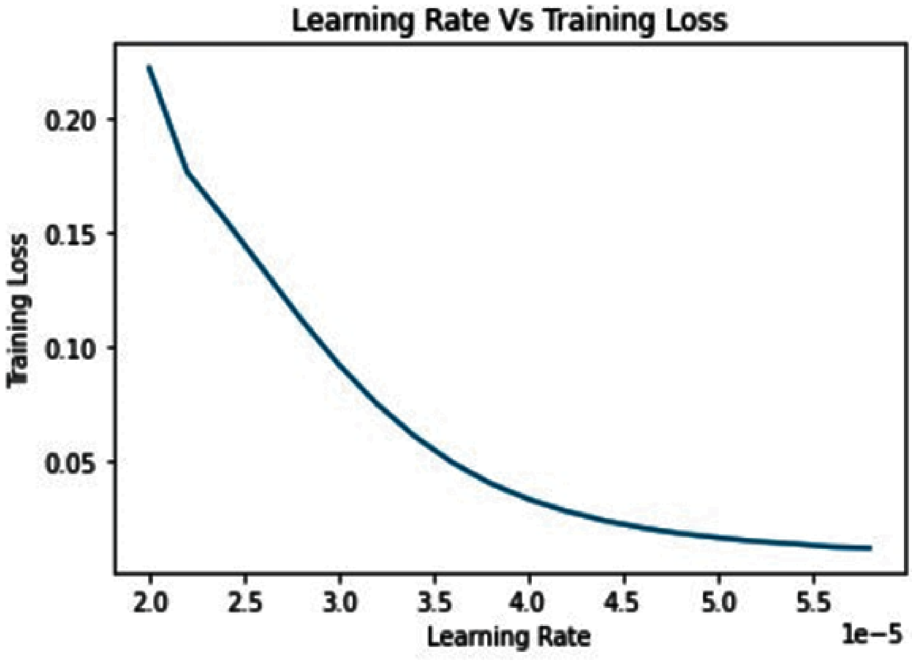


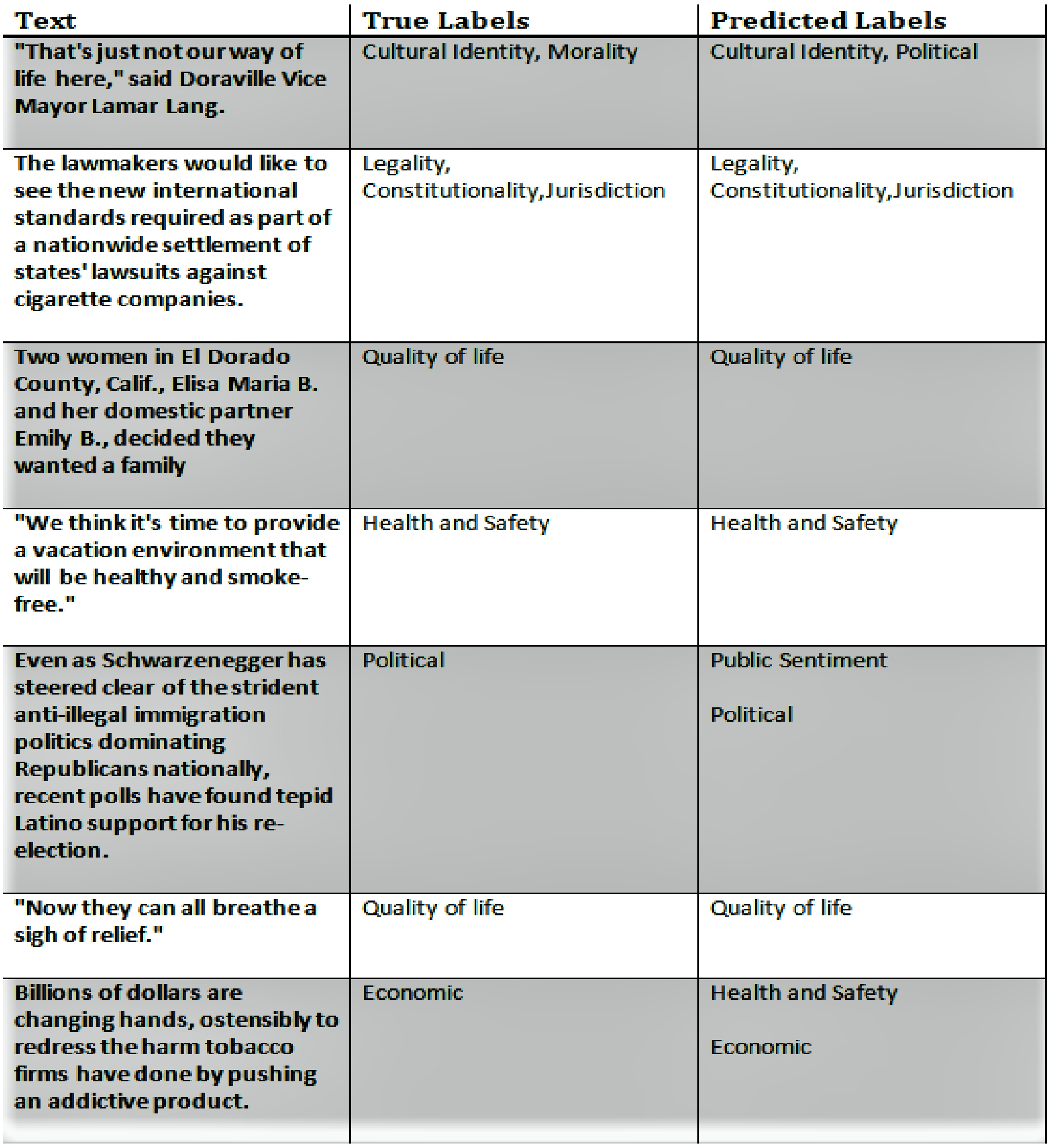
Figure 16: Illustration of training loss across a range of different learning rates. The lowest training loss is achieved at 5.8e-5

In this way, we set the training loss to 5.8e-5. By keeping the train-validation-test split as same, we trained the model with new hyper-parameters once again. In response to that, we achieved an accuracy of f1-micro: 0.63 and f1-macro: 0.57. For 12 out of 15 classes, we obtained an accuracy of 0.50 or higher. The result is shown by Table 4.

Apart from this quantitative evaluation of score metrics, we used our classifier to predict the frames of sentences extracted from online New York Times articles. We manually annotated the frames of those sentences and compared with the predicted ones. In most of the cases, the manual frames and predicted frames matched. Figure 17 describes one of the snapshots of this comparison:

Frame	Precision	Recall	F1 Score
Economic	0.72	0.69	0.71
Capacity & Resources	0.58	0.32	0.41
Morality	0.69	0.56	0.62
Fairness& Equality	0.65	0.50	0.56
Legality, Constitutionality & Jurisdiction	0.76	0.71	0.73
Policy Prescription & Evaluation	0.63	0.42	0.50
Crime & Punishment	0.70	0.63	0.66
Security & Defence	0.60	0.61	0.60
Health & Safety	0.76	0.57	0.65
Quality of Life	0.54	0.46	0.50
Cultural Identity	0.63	0.43	0.51
Public Sentiment	0.59	0.53	0.56
Political	0.76	0.73	0.74
External Regulation & Reputation	0.55	0.38	0.45
Other	0.50	0.34	0.40
Micro Average	0.69	0.59	0.63
Macro Average	0.64	0.53	0.57

Table 4: Representation of precision, recall and f1-score for each class (Frame).



Text	True Labels	Predicted Labels
"That's just not our way of life here," said Doraville Vice Mayor Lamar Lang.	Cultural Identity, Morality	Cultural Identity, Political
The lawmakers would like to see the new international standards required as part of a nationwide settlement of states' lawsuits against cigarette companies.	Legality, Constitutionality,Jurisdiction	Legality, Constitutionality,Jurisdiction
Two women in El Dorado County, Calif., Elisa Maria B. and her domestic partner Emily B., decided they wanted a family	Quality of life	Quality of life
"We think it's time to provide a vacation environment that will be healthy and smoke-free."	Health and Safety	Health and Safety
Even as Schwarzenegger has steered clear of the strident anti-illegal immigration politics dominating Republicans nationally, recent polls have found tepid Latino support for his re-election.	Political	Public Sentiment Political
"Now they can all breathe a sigh of relief."	Quality of life	Quality of life
Billions of dollars are changing hands, ostensibly to redress the harm tobacco firms have done by pushing an addictive product.	Economic	Health and Safety Economic

Figure 17: Comparison of manually annotated and predicted frames at sentence level. In the picture; Text refers to sentence from New York Times online articles, True Labels means manually annotated frames and Predicted Labels are frames predicted by the classifier.

In Figure 17, the first row of the table can be considered as an ambiguous sentence. The sentence talks about ’way of life here’. As per Figure 9, this phrase corresponds to Cultural Identity . As the sentence also had the word ’Mayor’, the classifier predicted both Cultural Identity and Political frames for the given sentence. Similarly, the last sentence of Figure 17 is about exchange of money to redress the deeds of tobacco companies. So we annotated this sentence as Economic. However, words such as ’harm’ and ’addictive’ led classifier to predict an additional frame as Health and Safety. Overall, in majority of cases, the classifier correctly performed at least one of the manually annotated frames. Thus, we conclude this classifier is good enough to use in further experiments of this thesis.

4.2.7 Comparing The Results of Classifiers at Different Text Levels

Based upon the performance results obtained in previous sections, we rank the sentence level classifier as first, article level as second and span level as last. Acknowledging the fact that both the classifiers at article level and sentence level used maximum sequence number as 128, we used sentence level classifier to classify articles and paragraphs in Chapter 5. The result of the classifiers at different levels is summarised in Table 5.

Level	Micro-f1	Macro-f1
Sentence	0.63	0.57
Article	0.60	0.48
Span	0.41	0.39

Table 5: Comparison of classifiers at different text levels. Level represents the text level.

4.2.8 Analysis of Frames and Persuasion Effect

Since Media Frames Corpus does not contain annotations for persuasion effect, we used Webis-Editorial-Quality-18 corpus consisting of 979 editorials. We used our frame classifier to identify frames and associate editorials to their respective frames. After that, we linked persuasiveness effect labels with their corresponding frames in order to study the relation between the two.

Figure 18 illustrates the distribution of effect labels across 15 frames for liberal. Likewise, Figure 19 represents the distribution for conservative.

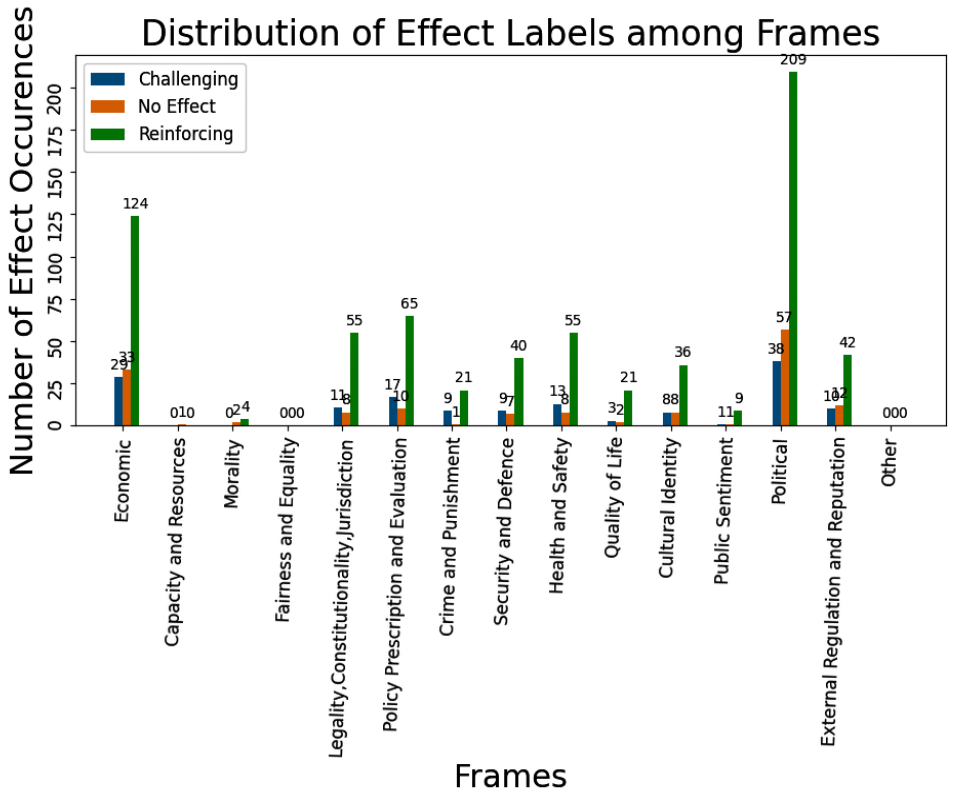


Figure 18: Number of occurrences of persuasiveness effect labels per frame in 979 editorials for liberal.

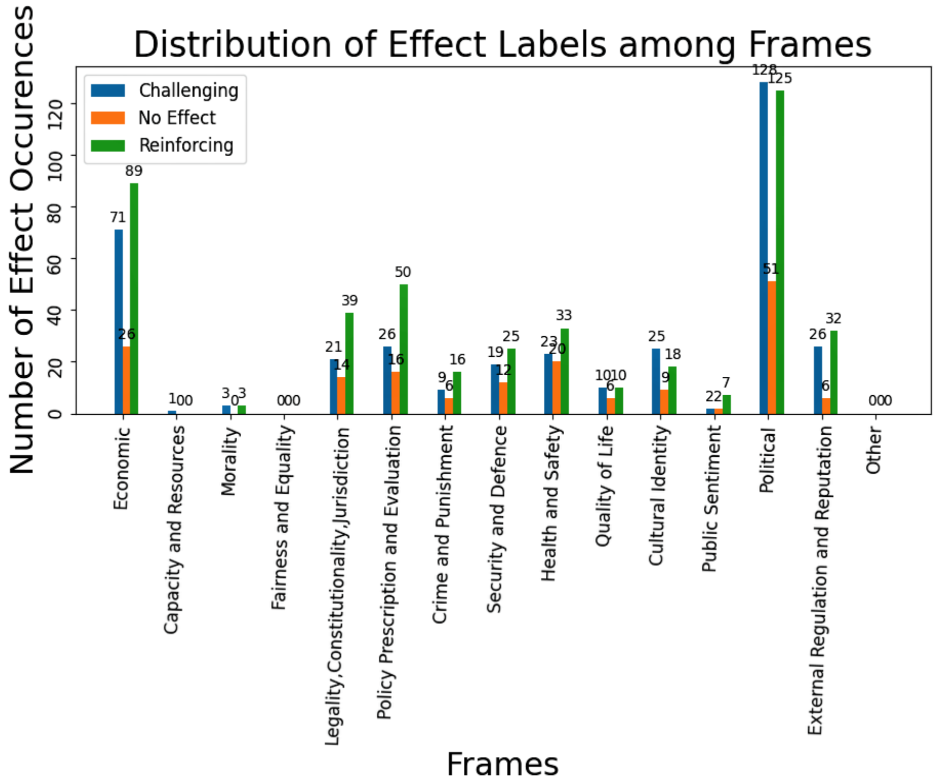


Figure 19: Number of occurrences of persuasiveness effect labels per frame in 979 editorials for conservative.

As we can observe in Figure 18 and 19, we found that frames such as Economic and Political were the most persuasive ones. Whereas, Fairness and Equality, Morality, Capacity and Resources and Other were sparsely used to persuade, if used at all. Moreover, a major contrasting difference between Figure 18 and Figure 19 (representing liberal and conservative respectively) is that the conservatives are more challenged by most of the frames as compared to liberals. Especially, frames like Political, Economic, and External Regulation and Reputation significantly challenge conservatives than liberals.

One of the reasons behind such low occurrences of some frames could also be the classifier sometimes missing to identify those frames (as discussed above in section 4.2.5). However, we acknowledge this fact and base our findings on overall general comparison.

5 Experiments and Results

In this section, we conduct experiments on the Webis-Editorial-Quality-18 corpus (El Baff et al.,2018) in order to enhance the prediction of persuasiveness and also to explore the effect of new content features (topic and frame). We use the existing style and content features from (El Baff et al.,2020) which are already described in Chapter 2. On top of that, we use our topic model and frame classifier to extract topics and frames respectively from the Webis-Editorial-Quality-18 corpus, and add them to the existing features as new content features. We do not add our sentiment feature because it is already included in other style feature as NRC EmotionSentiment.

5.1 Data

Webis-Editorial-Quality-18 corpus contains the annotations that distinguish if the editorials challenge the prior stance of the readers or reinforces their existing stance. As already defined before, 'challenging' refers to making annotators rethink their prior stance, but not necessarily change it. Whereas, 'reinforcing' means helping them argue better about a discussed topic(El Baff et al.,2018). Other editorials are annotated as 'no effect'.

Furthermore, each editorial was annotated by 3 liberal and 3 conservative annotators. The final persuasive effect is defined on the basis of majority vote of their annotations. As El Baff et al. 2020 found 21 duplicate editorials with the same content but different ids, the final number of editorials in the corpus is 979. These 979 editorials are chronologically split into oldest 80% as training set and newest 20% as test set. Figure 20 shows us the distribution of persuasive effect labels among training and testing sets in the corpus.

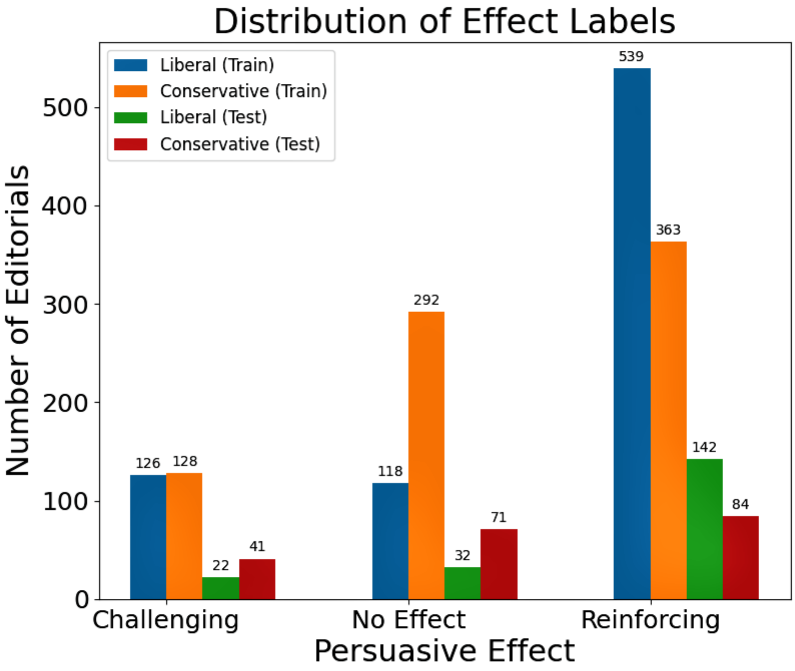


Figure 20: Distribution of persuasive effect labels among training and testing sets in the corpus.

5.2 Integration of Topics in the Corpus

In order to add Topics as content feature in the corpus, we trained the LDA model on the training set which includes 783 editorials. To define the optimal number of topics (k),we calculated the coherence score for k ranging between 10 to 100. We found that highest score (0.58) was for k=18. We associated those 18 topics with their corresponding 979 editorials in the corpus. The distribution of topics is displayed in Figure 21.

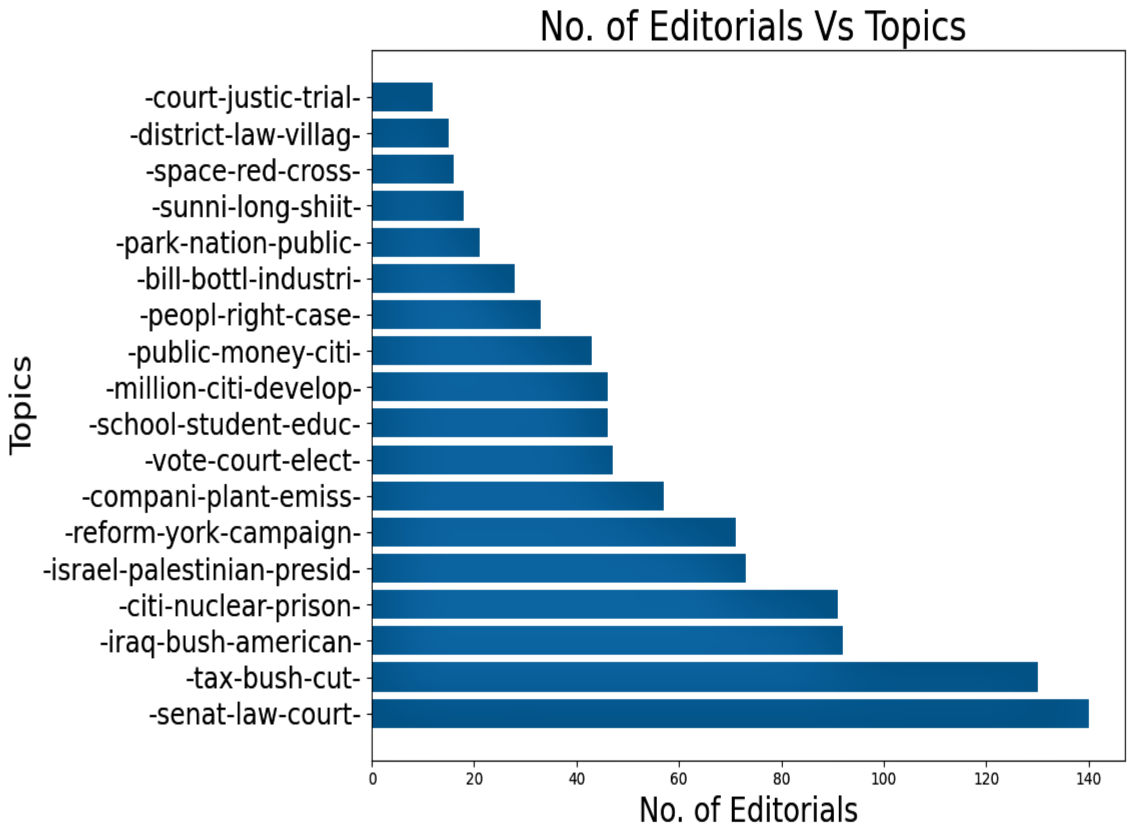


Figure 21: The distribution of 18 topics across 979 editorials in the corpus. Y-axis represent topics with their top 3 terms.

5.3 Integration of Frames in the Corpus

For this task, we used our pre-trained Frames classifier from Section 4.2 to classify 979 editorials into 15 frames. The distribution of obtained frames is showed in Figure 22.

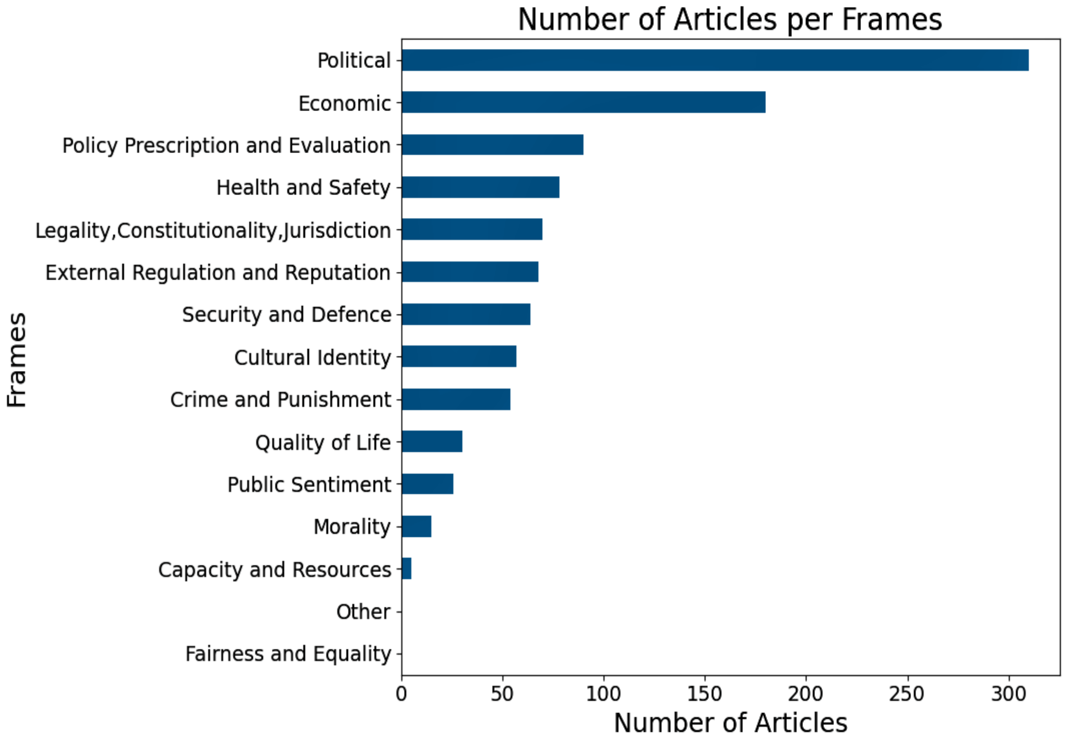


Figure 22: The distribution of 15 topics across 979 editorials in the corpus. The classifier did not assign any editorial to 'Fairness and Equality' and 'Other'.

5.4 Development of Classifier for Prediction of Persuasive Effects

Given the two ideologies: Liberal and Conservative, the prime goal of the classifier is to predict the persuasive effect of the editorials for each of the given ideologies. The input features for the classifier are the style features (LIWC, NRC EmotionSentiment, Webis ADUs, MPQA Arguing and MPQA Subjectivity) and content features (Lemma, Topics, Frames). The classifier must predict whether the editorial is challenging, reinforcing or ineffective (no effect).

Similar to work done by El Baff et al.2020, we trained one Support Vector Machine (SVM) model on the training set for each feature and their combinations as well. SVM is a supervised learning method that is used for classification, regression and also outliers detection. Apart from the advantages such as efficiency in high dimensional spaces, we used SVM in order to compare our prediction results with that of El Baff et al. 2020.

Next, we performed tuning of the model same as in El Baff et al.2020. We tuned our SVM model's (with linear kernel) cost hyper-parameters with the help of grid search. Grid search tunes the model by computing the optimum values of hyper-parameters. In order to provide grid search model with the opportunity to train on multiple train-test splits for a better indication of performance on unseen data, we used 5-fold cross validation method for grid search. We trained our SVM model using Scikit-learn (Pedregosa et al.,2011) library.

As Figure 23 shows us that the distribution of the effect labels is highly skewed, we assigned the hyper-parameter 'class_weight' to 'balanced'. Then, we trained the best model provided by SVM on the training set of 783 editorials, and tested on 196 editorials.

While using the combination of style and content (lemma) features, El Baff et al. 2020 got macro-f1: 0.43 and micro-f1: 0.54 for liberals, and macro-f1: 0.36 and micro-f1: 0.36 for conservatives. By training our SVM model with Topics and Frames features, we improved the performance (f1 scores) for conservatives by 0.01. On the other hand, the f1 score for liberals remained the same. Table 6 describes the summary of this result.

	Liberal			**Conservative**	
Feature	**Macro**	**Micro**	**Feature**	**Macro**	**Micro**
LIWC, MPQA Arguing, MPQA Subjectivity, Lemma (Content)*	0.43*	0.54*	MPQA Arguing, Lemma (Content)*	0.36*	0.36*
LIWC, MPQA Arguing, MPQA Subjectivity, Lemma (Content)'	0.43'	0.54'	LIWC, Webis ADUs, Lemma (Content), Frame (Content)'	0.37'	0.38'

Table 6: Evaluation metrics (micro and macro F1-scores) of the best content + style combinations in classifying the persuasive effect on liberals and conservatives. Features and scores with * refers to El Baff et al. 2020 and ' refers to our work.

In order to further improve prediction, we introduced a new layer of feature selection on top of the existing one. Apart from getting the best combination and using only the features of that combination, we used univariate selection on the features. Univariate feature selection chooses the best features based on univariate statistical tests. For this purpose, we used 'SelectKBest' class from scikit-learn (Pedregosa et al.,2011).

SelectKBest selects k highest scoring features and removes others, where k is the input that asks for desired number of features to keep in the model. The score is based on the chi-squared statistical test that assess the relation between two categorical features where higher score means more valuable relation. In our case, we evaluated the relation between each feature and our target variable (persuasive effect labels). As a result, we got features that were important for the model to predict well, and got rid of those which were not important and hampered the quality of prediction.

Selection of appropriate number of features for the model was crucial in order to hold all the important features while not letting any less valuable feature in the feature space. So, we checked the performance of the model with a series of different numbers of features (k) and selected the k which scored the highest. We obtained highest score while using k=50 for liberal(shown in Table 10) and conservative (shown in Table 11). For both ideologies, the SelectKBest scores for all the selected features were approximately 3. Therefore, we keep the score of 3 as a threshold for selecting features in further experiments. An instance of such system of scoring is illustrated in Table 10. We visualise the performance of the model across different k in Figure 23.

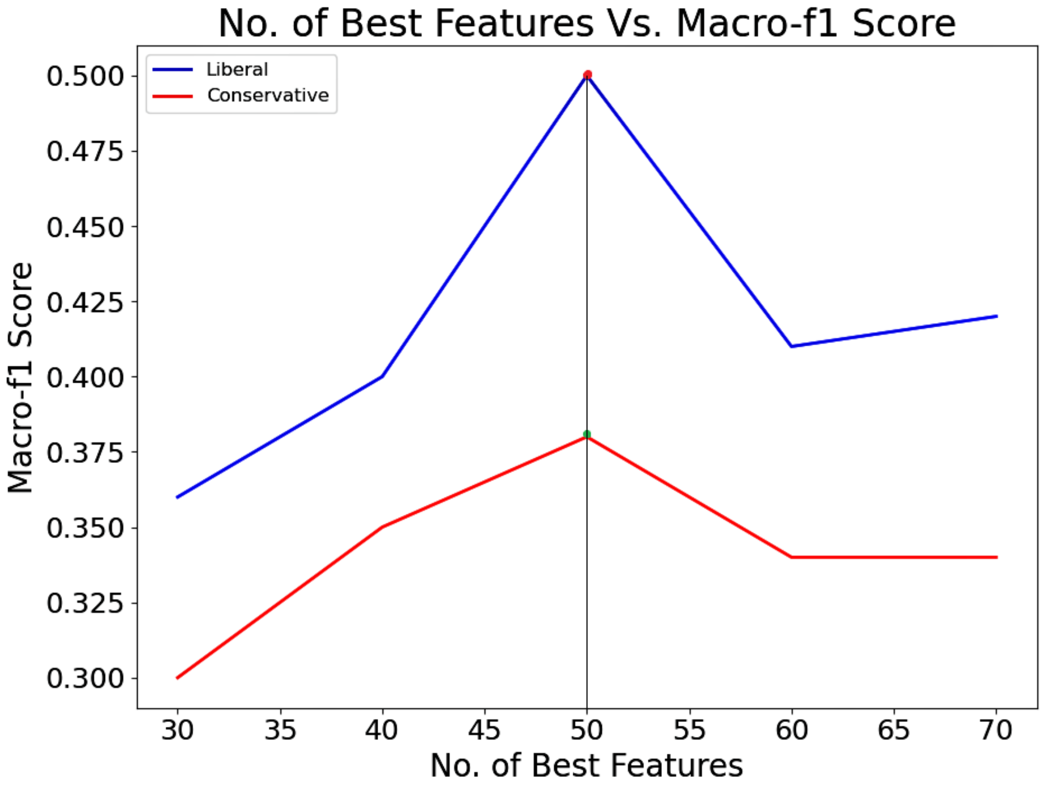


Figure 23: The performance of the model among a series of k ranging from 30 to 70. K=50 yields the best macro-f1 score for both liberal and conservative.

Finally, we ran our new SVM model with the same train and test sets. Table 7 illustrates the scores of individual features and the best combination for liberal and conservative. For liberal, the best combination contains features such as MPQA Arguing, Lemma and Topic. Similarly, for conservative, the best combination is comprised of Lemma and Frame.

There are five pivotal things to notice in Table 7. They are as follows:

	Liberal		Conservative	
Feature	**Macro**	**Micro**	**Macro**	**Micro**
LIWC	0.32	0.43	0.32	0.33
NRC Emotion&Sentiment	0.34	0.44	0.31	0.32
MPQA Arguing	0.19	0.19	0.24	0.27
MPQA Subjectivity	0.33	0.38	-	-
Lemma	0.48	0.61	0.38	0.44
Topic	0.33	0.57	-	-
Frame	0.07	0.11	0.24	0.31
Top features by SelectKBest	**0.50**	**0.61**	**0.42**	**0.46**

Table 7: Evaluation metrics (micro and macro F1-scores) of each feature type and their best combination in classifying the persuasive effect on liberals and conservatives. The scores in **bold** represent the highest scores obtained. (**-**) in place of score means the feature could not make it to top 50 features while feature selection.

1 Webis ADUs was not selected as top features due to its weaker relation (as compared to other features) with persuasion effect labels.

2 Just like Webis ADUs, Topic is also not included for conservative because it was not within the top 50 features. This suggests that Topic is less relevant while predicting persuasiveness effect in case of conservative. However, it is included in the best feature combination for liberal. This means that Topic is more important (in prediction of persuasiveness effect) for liberal than conservative.

3 The f1 scores of Frame for liberal is low. However, it has higher score in case of conservative. Plus, it is one of the feature among best feature combination for conservative. It also means that Frame is more important (in prediction of persuasiveness effect) for conservative than liberal.

4 We significantly increased the f1 scores for both liberal and conservative (by 0.07 and 0.05 respectively) as compared to our previous prediction model.

5 In comparision to El Baff et al. 2020, we increased the f1 scores by 0.07 for liberal and by 0.06 for conservative.

5.5 Addition of Features from Editorials' Paragraphs

The main objective of this experiment is to determine whether adding the style and content features of paragraphs to our existing features in article level result in better prediction scores. The experiment is divided into two phases where we train models with and without feature selection. The reason behind such approach is to analyse whether the model predicts better with all the features or with just the best features and their combination.

Apart from annotations in article level, Webis-Editorial-Quality-18 corpus also contains annotations in paragraph level for the same editorials. It includes same style features as in article level but lacks content features. Therefore, we added topics, frames and lemma as content features.

In order to classify paragraphs into different topics, we used our LDA model from Section 4.4.2 and associated each paragraph to its corresponding topic. Similarly, we also used our frame classifier from Section 4.4.3 to classify paragraphs into 15 frames.

The crucial part of this experiment was to define the paragraph features. Since, the distribution of paragraphs across editorials was uneven, it was pivotal to select the same number of paragraphs from each editorial. So, we defined our paragraph level features based on four scenarios. They are as follows:

1 **First and last paragraphs** We believe that first (introduction) paragraph and last (conclusion) paragraph are important parts of an editorial. The first paragraph usually includes bits and pieces of information regarding the title of the editorial (Rich,2015). On the other hand, the last paragraph usually summarises the main points of the editorial(Cla,2009). Therefore, we extracted the features of first and last paragraphs and added them with the features of article (aka editorial). With the help of common 'editorial id' we linked and stored these features with the features of their respective editorials.

   Finally, we trained the model once with feature selection (selecting only top features using SelectKBest and choosing best combination from those features), and again without feature using all features.

   With feature selection, we obtained accuracy of macro-f1: 0.44 and micro-f1: 0.56 (for liberal) and macro-f1: 0.45 and micro-f1: 0.45 (for conservative). Whereas, without feature selection, the classifier scored macro-f1: 0.37 and micro-f1: 0.53 (for liberal) and macro-f1: 0.34 and micro-f1: 0.35.

   In case of feature selection, the best combination including features from first and last paragraphs are presented in Table 9. Since, the obtained f1 scores are lower than the best score in both Table 7 and Table 9, we conclude that adding first and last paragraphs' features did not improve the classifier.

2 **Paragraphs having same topic/frame as that of its respective editorial**

   A paragraph is certainly more related to the editorial if it has the same topic or frame or both. We extracted those type of paragraphs from all the editorials in the corpus and added to the existing article level features.

Unfortunately, we found only 184 editorials with at least one paragraph having same topic as the article. However, we got 916 editorials with at least one paragraph having same frame as the article.

We did not consider adding the topic feature due to its low number of examples. On the other hand, 916 out of 979 editorials was a pretty decent number for training. So, we extracted all the paragraphs from their respective editorials that had common frames. Then, we calculated the average number for each feature (e.g. NRC_emotion sad) for all paragraphs (with same frame) per article. The rest of the examples referring to 63 unselected editorials were filled with zeros for numerical features and with none for categorical features.

With the same train and test sets, we ran the model in the same way as before. With SelectKBest features plus best combination, we obtained accuracy of macro-f1: 0.47 and micro-f1: 0.60 (for liberal) and macro-f1: 0.41 and micro-f1: 0.43 (for conservative). Whereas, without feature selection, the classifier scored macro-f1: 0.28 and micro-f1: 0.39 (for liberal) and macro-f1: 0.29 and micro-f1: 0.29.

The f1 scores for this one are also lower than the best score in Table 8. Therefore, we can say that adding features of the common frame paragraphs did not improve the classifier as well.

3 **Paragraph with highest ranking Tf-Idf words.**

Tf-Idf evaluates how relevant a word is to a document in a collection of documents. It depends on the number of times a word occurs in a document, and the inverse document frequency of the word across a set of documents. We used this statistical measure to find paragraph containing most relevant word in the editorials, and added the paragraphs' features with the features of article level.

As a result of this new dataset, we obtained accuracy of macro-f1: 0.48 and micro-f1: 0.58 (for liberal) and macro-f1: 0.43 and micro-f1: 0.43 (for conservative) with feature selection. Whereas, without feature selection, the classifier scored macro-f1: 0.33 and micro-f1: 0.43 (for liberal) and macro-f1: 0.31 and micro-f1: 0.32. Much like previous experiments, this experiment also did not yield a better result.

4 **Paragraphs with question mark(?) and/or quotation mark("")**

The main idea behind this sort of selection was based on our observation on some of the editorials. We noticed that editorials use rhetorical questions to empower their arguments. Moreover, they also quote statements from different personalities in order to prove their arguments' authenticity. Thus, from this standpoint, we extracted such paragraphs with question and/or quotation marks.

> In contrast to our expectation, we found only 292 editorials that had at least one paragraph containing either quotation or question marks. Therefore, we dropped this experiment due to insufficient number of examples.

As our last experiment, we added features from points 1, 2 and 3 above with the article level features. By doing so, we got accuracy of macro-f1: 0.36 and micro-f1: 0.53 (for liberal) and macro-f1: 0.45 and micro-f1: 0.46 (for conservative) with feature selection. Whereas, without feature selection, the classifier scored macro-f1: 0.35 and micro-f1: 0.51 (for liberal) and macro-f1: 0.34 and micro-f1: 0.36.

The results are summarised below in Table 8 and 9.

Using All Features			
Features of Article(A), Paragraph(P)	**Ideology**	**Macro-f1**	**Micro-f1**
A	Liberal	**0.43**	**0.56**
	Conservative	0.30	0.31
A + First and Last P	Liberal	0.37	0.53
	Conservative	0.34	0.35
A + P with Highest Tf-Idf	Liberal	0.33	0.43
	Conservative	0.31	0.32
A + P with Matching Frame	Liberal	0.28	0.39
	Conservative	0.29	0.29
A + First and Last P **+ P with Highest Tf-Idf** **+ P with Matching Frame**	Liberal Conservative	0.35 **0.34**	0.51 **0.36**

Table 8: Evaluation metrics (micro and macro f1 scores) of each features at article and paragraph levels in classifying the persuasive effect on liberals and conservatives. The scores in **bold** represent the highest scores obtained.

5.6 Discussion of Results

For liberal, we got the best f1 score from the model trained at article level with SelectKBest features and their best combination. We obtained macro-f1: 0.50 and micro-f1: 0.61. These scores surpass the best scores in El Baff et al. 2020 and all of our other models.

The best combination for liberal contained MPQA Arguing, Lemma and Topic. This suggests that different patterns of argument like authority, assessments, emphasis and doubt when combined with topics and lemma yields the best prediction of persuasiveness effect for liberal. Both Lemma and MPQA Arguing were present in the best combination in El Baff et al. 2020. So, Topic being the newly added feature was definitely associated with the improvement in prediction.

On the other hand, for conservative, we achieved the best score (article level) of macro-f1: 0.44 and micro-f1: 0.46 with the best combination: Lemma and Frame. This score also exceeded the best score in El Baff et al. 2020 where frame was not used. So, we can also say that Frame has a role to play in improving the prediction results.

Whether we talk about our results or the result of El Baff et al. 2020, Lemma was consistently found in the best combination for conservative. Therefore, it can be considered as one of the most valuable feature for prediction of persuasion.

Moreover, when we combined the features of article and paragraphs (first and last paragraphs, paragraph with word containing highest tf-idf and paragraph with matching frame) for conservative, the best score further improved to macro-f1: 0.45 and micro-f1: 0.46. The best combination included NRC EmotionSentiment (article level feature), Lemma (article level feature), LIWC (feature from first paragraph), MPQA Arguing (feature from first paragraph), LIWC (feature from last paragraph), NRC Emotion-Sentiment (feature from last paragraph) and LIWC (feature from paragraphs with matching frame). We found that LIWC was an important style feature for conservative which occurred thrice from different text levels in our best combination.

The combination of article and paragraph level features could not exceed the best score (of article level). We obtained macro-f1: 0.48 and micro-f1: 0.58 as the highest score among the models with combined article and paragraph level features. We got these scores when we combined features of paragraph with highest tf-idf scoring word, and article. But still, highlight of the result was the presence of Topic and Frame in the best combination. The best combination was: LIWC (article), Topic (first paragraph), MPQA Subjectivity (Article), Lemma (Article), MPQA Arguing (last paragraph), Frame (paragraph with highest tf-idf) and Topic (paragraph with highest tf-idf).

Using Feature Selection				
Features of Article(A), Paragraph (P)	**Ideology**	**Macro-f1**	**Micro-f1**	**Best Combination**
A	Liberal	**0.50**	**0.61**	MPQA Arguing, Lemma Topic
	Conservative	0.42	0.46	Lemma, Frame
A + First and Last P	Liberal	0.44	0.56	art_Lemma, fpar_MPQA Arguing, fpar_Frame
	Conservative	0.45	0.45	art_MPQA Arguing, art_Lemma, fpar_LIWC, fpar_MPQA Arguing, lpar_LIWC, lpar_MPQA Arguing lpar_MPQA Parargaph Subjectivity, lpar_Frame, lpar_Topic
A + P with Highest Tf-Idf	Liberal	0.48	0.58	art_liwc, art_MPQA Subjectivity, art_Lemma, art_Topic, tfpar_MPQA Arguing, tfpar_Frame, tfpar_Topic
	Conservative	0.43	0.43	art_LIWC, art_MPQA Arguing, art_Lemma, tfpar_MPQA Arguing
A + P with Matching Frame	Liberal	0.47	0.60	art_Lemma
	Conservative	0.41	0.43	art_MPQA Arguing art_Lemma, fpar_NRC
A + First and Last P + P with Highest Tf-Idf + P with Matching Frame	Liberal	0.36	0.53	fpar_Topic, lpar_MPQA Arguing, tfpar_NRC, tfpar_Frame, mfpar_LIWC, mfpar_NRC
	Conservative	**0.45**	**0.46**	art_NRC, art_Lemma, fpar_LIWC, fpar_MPQA Arguing, lpar_LIWC, lpar_Nrc, mfpar_LIWC

Table 9: Prediction accuracy scores of models with article and paragraph level features and their best combination selected by feature selection. The scores in **bold** represent the highest scores obtained. art_ refers to article level, fpar_ refers to first paragraph, lpar_ refers to last paragraph, tfpar_ refers to paragraph with highest tf-idf and mfpar_ refers to paragraph with matching frame.

6 Conclusion

In this thesis, we study the importance of content features such as topic and frame combined with style features in news editorials for persuading readers with different political stances.

We perform topic modelling on editorials to capture the information on what an editorial is about. We also develop media frame classifier and recognise how editorials are framed for persuasion. By using the existing style features, topic and frame, we design a system that can predict the effect of persuasive editorials.

From our findings, we show that persuasion does not just depend on how the language is used. It also relies upon what the conveyed message is about and what specifically filtered information it contains.

We unfold the topics and frames that play significant role in persuading the readers of editorials, revealing how some of the topics as well as frames are more abstract than the others which becomes a challenge for the classifying model.

Acknowledging the significance of certain kinds of paragraphs in editorials, we implement the features of the paragraphs combined with that of articles, and notably improve the prediction of persuasion for one of the political ideologies.

Besides our findings and achievements, there is still room for further improvements. For instance, the detection of media frames can be further investigated. More sophisticated approaches can be developed to prioritise the important keywords related to some of the ’abstract’ frames that our classifier found difficult to detect.

Moreover, one can try other new methods of combining article level and paragraph level features. An extensive experiment can consider the combination of paragraph level features integrated with features of sentences. To that end, creation of sentence level corpus annotated with appropriate style and content features can be one of the future tasks to perform.

Furthermore, we focused on political ideology of the readers in this thesis. The personality traits of the readers can also be considered to study persuasion in editorials. In other words, instead of analysing persuasion on liberal and conservative, persuasion can be studied on different personality types based on various personality traits. With personality traits, the scope of the study can be significantly widened from political groups to individual people. In this way, our work can also be adapted for further applications in several domains such as persuasive essays, online persuasive blogs and so on.

List of Figures

List of Tables

Bibliography

[1] Understanding latent dirichlet allocation (lda). https:/www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/.

[2] Dbpedia spotlight. https:/www.dbpedia-spotlight.org.

[3] Intuitive guide to latent dirichlet allocation. https:/towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158/.

[4] Pew research center. https://www.pewresearch.org/politics/quiz/political-typology/.

[5] The writing center. https://www.clarion.edu/academics/student-success-center/writing-center/67205.pdf, 2009.

[6] Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Pro-*

ceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3433–3443, 2016.

[7] Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, 2017.

[8] Amparo Elizabeth Cano Basave and Yulan He. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, 2016.

[9] Amber E Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A Smith. Tracking the development of media frames within and across policy issues. 2014.

[10] Katarzyna Budzynska and Chris Reed. Advances in argument mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 39–42, 2019.

[11] Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, 2012.

[12] Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.

[13] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological science*, 15 (10):687–693, 2004.

[14] Michael Scott Davis. *Editorial personality: Factors that make editorial writers successful.* PhD thesis, University of Missouri–Columbia, 2013.

[15] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, 2018.

[16] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Analyzing the persuasive effect of style in news editorial argumentation. Association for Computational Linguistics, 2020.

[17] Lewis R Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

[18] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, 2016.

[19] Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4 (1):89–106, 2016.

[20] Ralph Henry Johnson and J Anthony Blair. *Logical self-defense*. Idea, 2006.

[21] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014.

[22] Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *12th ACM Conference on Web Science*, pages 305–314, 2020.

[23] Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*, 2017.

[24] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[25] Ana Laura Nettel and Georges Roque. Persuasive argumentation versus manipulation. *Argumentation*, 26(1):55–69, 2012.

[26] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[28] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844, 2014.

[29] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[30] Richard E Petty, John T Cacioppo, and Rachel Goldman. Personal involvement as a determinant of argument-based persuasion. *Journal of personality and social psychology*, 41(5):847, 1981.

[31] Radim Řehřek, Petr Sojka, et al. Gensim—statistical semantics in python. *Retrieved from genism. org*, 2011.

[32] Carole Rich. *Writing and reporting news: A coaching method*. Cengage Learning, 2015.

[33] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112, 2003.

[34] Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34, 2007.

[35] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[36] Teun A Van Dijk. Opinions and ideologies in editorials. In *4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought, Athens*, pages 14–16, 1995.

[37] Tuija Virtanen and Helena Halmari. Persuasion across genres. *Persuasion across genres: A linguistic approach*, 130:3, 2005.

[38] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, 2017.

[39] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, 2017.

[40] Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232, 2017.

[41] Wikipedia. Topic model. `https://en.wikipedia.org/wiki/Topic_model/`.

[42] Ladislav Zgusta. Lexicography then and now. In *Lexicography Then and Now*. Max Niemeyer Verlag, 2012.

7 Appendix

Top 50 Features (Liberal)	
Feature	**Score**
lemma3_presid mahmoud abba	2.926952
mpqa_arg_wants	2.933042
lemma3_suffer mental ill	2.986976
lemma3_squar foot	3.005193
lemma2_counti execut	3.214585
lemma2_poor countri	3.252541
lemma1_congress	3.268466
lemma2_suffolk counti	3.302473
lemma3_weapon mass destruct	3.320892
nrc_anticipation	3.321942
lemma3_hundr million dollar	3.388733
lemma2_long island	3.411844
mpqa_arg_priority	3.430544
frame	3.505158
lemma3_world trade center	3.673057
lemma3_major leader joseph	3.701273
lemma3_mayor michael bloomberg	3.760999
lemma3_york state capitol	3.764981
lemma1_governor	3.825794
lemma1_campaign	3.845404
mpqa_arg_structure	3.975651
lemma2_global warm	4.079103
lemma1_school	4.143133
nrc_trust	4.293905
lemma1_citi	4.513139
lemma2_york citi	4.557340
lemma2_eliot spitzer	4.606756
lemma1_reform	4.668436
lemma2_re elect	4.766273
mpqa_subjobg_obj	4.798559
lemma2_state senat	5.086245
mpqa_arg_contrast	5.362443
lemma3_unit nation secur	5.368268
lemma3_nation secur council	5.368268
lemma3_like york citi	5.735080
lemma2_campaign financ	6.494374
lemma3_state suprem court	7.189805
mpqa_subjobg_subj	7.743849
nrc_disgust	8.076099
nrc_anger	8.413418
liwc_scores_clout	8.850848
nrc_joy	11.864984
liwc_scores_wc	16.179398
nrc_sadness	19.758349
topic	25.870114
nrc_negative	31.684697
nrc_positive	32.288338
nrc_fear	41.016492
liwc_scores_tone	214.202353

Table 10: Top 50 features for liberal along with their univariate selection score.

Top 50 Features (Conservative)	
Feature	**Score**
lemma2_campaign financ	2.683719
liwc_scores_wps	2.719114
lemma2_fund rais	2.806570
lemma2_million peopl	2.857712
lemma3_prime minist ariel	2.908181
lemma3_minist ariel sharon	2.908181
lemma2_york citi	2.929465
lemma2_look like	2.950351
lemma3_senat nichola spano	2.959540
lemma2_sunni arab	3.050095
lemma3_port author york	3.109427
lemma2_trade center	3.183419
lemma3_health human servic	3.220280
lemma3_arm servic committe	3.384245
lemma2_west bank	3.528249
lemma2_suprem court	3.603678
lemma3_good govern group	3.623261
nrc_anticipation	3.718903
nrc_surprise	3.773817
lemma2_million year	3.909193
lemma3_attorney general offic	4.281877
nrc_joy	4.559776
lemma2_north korea	4.690423
lemma2_feder govern	4.755126
lemma3_suprem court decis	4.892208
lemma3_suffer mental ill	4.913145
lemma2_hedg fund	5.060726
lemma3_nuclear weapon program	5.122406
mpqa_arg_rhetoricalquestion	5.361251
mpqa_arg_generalization	5.381455
lemma3_attorney general eliot	5.443847
lemma3_general eliot spitzer	5.443847
lemma2_illeg immigr	5.972278
lemma3_world trade center	5.973593
mpqa_arg_priority	6.036167
mpqa_arg_emphasis	6.097251
lemma3_fuel effici car	6.790119
lemma3_hundr million dollar	6.806046
lemma3_senat john mccain	6.861113
lemma2_billion dollar	6.905488
lemma2_state attorney	7.278714
liwc_scores_wc	9.010656
frame	9.925168
nrc_positive	10.598857
nrc_sadness	12.857295
nrc_negative	13.229637
nrc_disgust	17.202535
nrc_anger	24.457566
nrc_fear	26.987198
liwc_scores_tone	160.178885

Table 11: Top 50 features for conservative along with their univariate selection score.