Modelling Argumentation in Short Text: a Case of Social Media Debate

Anastasios Lytos[1], Thomas Lagkas[2], Panagiotis Sarigiannidis[3], Vasileios Argyriou[4], and George Eleftherakis[5]

[1] Department of Computer Science, University of Sheffield, Sheffield, UK
alytos1@sheffield.ac.uk

[2] Department of Computer Science, International Hellenic University, Kavala Campus, Greece
tlagkas@cs.ihu.gr

[3] Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece
psarigiannidis@uowm.gr

[4] Department of Networks and Digital Media, Kingston University, Kingston upon Thames, United Kingdom
vasileios.argyriou@kingston.ac.uk

[5] CITY College, University of York Europe Campus, Thessaloniki, Greece
eleftherakis@york.citycollege.eu

Abstract. The technological leaps of artificial intelligence (AI) and the rise of machine learning have triggered significant progress in a plethora of natural language processing (NLP) and natural language understanding tasks. One of these tasks is argumentation mining which has received significant interest in recent years and is regarded as a key domain for future decision-making systems, behavior modelling, and natural language understanding problems. Until recently, natural language modelling tasks, such as computational argumentation schemes, were often tested in controlled environments, such as persuasive essays, reducing unexpected behaviours that could occur in real-life settings, like a public debate on social media. Additionally, the growing demand for enhancing the trust and the explainability of the AI services has dictated the design and adoption of modelling schemes to increase the confidence in the outcomes of the AI solutions. This paper attempts to explore modelling argumentation in short text and proposes a novel framework for argument detection under the name Abstract Framework for Argument Detection (AFAD). Moreover, different proof-of-concept implementations are provided to examine the applicability of the proposed framework to very short text developing a rule-based mechanism and compare the results with data-driven solutions. Eventually, a combination of the deployed methods is applied increasing the correct predictions in the minority class on an imbalanced dataset. The findings suggest that the modelling process provides solid grounds for technical research while the hybrid solutions that combine symbolic AI and data-driven approaches have the potential to be applied to a wide range of NLP-related tasks offering a deeper understanding of human language, reasoning, and behavior.

1 Introduction

The recent advances in Machine Learning (ML) and the easily accessible processing power via cloud computing services have triggered a series of changes in different computational fields demonstrating significant research and market potential. The inter-disciplinary field of Natural Language Processing (NLP) has attracted significant interest from the research community [25,?] in both theoretical and practical level. NLP-related tasks include language modelling, opinion mining, sentiment analysis, and information retrieval while data-driven approaches, such as deep learning, are usually adopted and deployed. However, the performance of the data-driven approaches for these tasks seems to have reached a plateau due to hardware limitations and the enormous amount of data they require [16]. The research questions that have emerged are now more complex and require high levels of intelligence and an in-depth understanding of the human language. The proposed frameworks and systems should be capable of understanding not only what people think, but also identify the underlying reasons behind their stance on a given issue.

This question has prompted research in the field of argumentation mining, which is defined as a series of actions that could be independent or interconnected and they are relevant to the tasks of detection, extraction and evaluation of arguments [19]. The initial approaches on modelling arguments aimed at identifying a flawless argument in specific fields (Law, Scientific Papers) and serving specific needs (completeness, effectiveness) [26,?]. The developments beyond Web 2.0 have transformed the means of communication and information exchange, promoting shorter bursts of text without solid argumentation, while social networks have change the way information is shared [21,?]. In this noisy environment, it is important to develop mechanisms that are capable of identifying argumentation in short text revealing previously unexplored capabilities in the wider spectrum of the NLP domain.

The foundations of argumentation modelling lie on knowledge- and logic-based approaches [11,?], however in the last decade the majority of the research community prefers to adopt data-driven solutions [1,?,?], mostly due to the recent advances in Artificial Intelligence (AI). However, for a variety of problems in the wider area of NLP, solely data-driven solutions seem to have reached the upper boundaries of their potential [?,?] searching to advance background knowledge [18,?]. At the same time, rule-based approaches heavily rely on domain knowledge thus making knowledge transfer to different domains a difficult task [27,?]. Designing systems that can deploy context-aware mechanisms, such as collaborative filtering [2,?] or exploit any available metadata [28,?] is an increasingly important need.

Inspired from the need for modelling argumentation to enhance the trust and explainability in NLP systems, a series of definitions (topic, claim, reason, etc.) are provided, and the abstract framework for argumentation detection (AFAD) is proposed. Different implementations of the AFAD are presented, testing its performance and suitability in modelling short text. Additionally, a data-driven approach is also examined by implementing four ML algorithms. Finally, based on the idea of incorporating deterministic rules to probabilistic classifiers [9,?] a hybrid argumentation mining system is proposed. The predictions of the ML algorithms are calibrated through a rule-based mechanism measuring the semantic similarity between previously unseen chunks of text and a collection of arguments. The challenge is not only the transformation of the text into a binary format that a machine can understand but also the calculation of the similarity in the constructed mathematical representations.

The emerge of Web 2.0, and in particular social media, has provided tremendous potential in the NLP research community as it offers an endless and free source of data covering a wide spectrum of topics. On the other hand, it also sets a series of challenges for its successful modelling because of its often low quality content. When researchers choose social media as their data source, they tend to prefer political and social debates such as Brexit, national elections, constitution changes, etc. This work focuses on the construction of the natural gas pipeline *Nordstream2*, a topic that apart from its economic impact, has also emerged as a political debate in the European Union (EU) [14].

The main contribution of the paper can be summarised in five points:

- Presents a computational definition of argumentation in (very) short text and introduces the Abstract Framework for Argument Detection (AFAD)
- Demonstrates different proof-of-concept implementations for the provided definition of argumentation.
- Confirms the added value of integrating rule-based mechanisms into ML algorithms.
- Evaluates the performance of four different ML algorithms and the effect of applying additional features on them.
- Provides a publicly available annotated dataset for the existence of argumentation in short text.

The rest of the document is organized as follows: Section 2 discusses relevant literature review on argumentation modelling focusing on argumentation detection. Section 3 presents the motivation for argumentation detection in short text and provides the definition for the existence of an argument. Section 4 illustrates the process for modelling argumentation detection presenting a series of definitions and introduces the AFAD. Section 5 presents the experiments that took place including the description for the dataset collection, the implementation of the rule-based approach, the deployment of the ML algorithms, and the proposed hybrid solution. Section 6 presents the results of the experiments. Section 7 initiates a discussion on the findings of the research and presents the challenges. Finally, Section 8 draws conclusions and puts forward suggestions for future work.

2 Related Work

The wider field of argumentation mining has been flourishing in the recent years including both research in the field per se [12,?], but also expressing the interdisciplinary nature of the field through the development of applications in the legal domain [10] or adopting a cross-lingual approach [13,?]. The research of argumentation modelling schemes reveals the roots of the field [23,?], whereas data-driven analysis highlights relevant applications and tasks under the term argumentation mining [8,?]. The evolution of the field throughout the years and the impact of social media on the domain is also an interesting perspective [19].

This research paper bridges the two main schools of thought: informal logic and data-driven solutions. The contribution in the former lies on the definition of terms and theorems with a special focus on argumentation in very short text without proceeding in the exhaustive formal verification and without considering the chunks of text per-se argumentative [11,9]. The provided definitions and concepts can be re-used and applied on different domains. On the other hand, data-driven research on the task of argument detection usually takes place under a wider argumentation mining pipeline having as an endpoint either an evidence/source identification [1,?] or the visualization of the findings through argumentation graphs [6]. The datasets that are usually used in related work are either balanced or the positive instances outnumber the negative ones. On the contrary, the dataset we work with is strongly imbalanced towards the negative class creating a hostile environment for prediction mechanisms which, however, represents a real-life scenario.

Regarding hybrid, or close to hybrid approaches, an interesting combination of encoding techniques and supervised ML algorithms estimates the matching degree between a topic with known associated key points and previously unseen arguments [4]. However, the two approaches are not combined, but the word-embedding techniques feed a neural network algorithm that has been deployed on a later step. The open-loop architecture allowing an expert to interfere with the output of the ML algorithms has also been deployed, presenting promising results [10].

3 Motivation

Argumentation, as any multidisciplinary field, is open to different definitions and interpretations depending on the scope of each research study. Therefore, the motivation behind the different argumentation modelling attempts should be taken into consideration. In this work, argument is defined as a series of statements expressed in natural language, called premises, intended to support, and eventually determine the effectiveness of another statement, the conclusion. The above definition, although solid and complete, is elaborated and modelled in this section to fit the requirements of logic, computational argumentation and argumentation mining.

An argument's computational model can automatically assign strength values to a statement and evaluate various aspects of argumentation such as persuasion, cohesion, and stance detection. The success of the *abstract argumentation framework* [11] is merely due to its capability to get extended and modified to assign numerical values to a statement and eventually offer a natural link to statistical methods and tools. This ability of a computational model to quantify a concept that has qualitative substance is of major importance, because it could provide a crucial incentive to the development of human-level reasoning machines capable to interpret argumentation. Focusing on the field of argumentation mining in the context of social media, an explicit definition of argumentation (or argument) has not been given, as different scholars adopt different views and avoid providing a strict definition of argument.

Different norms and deduction processes are followed in social media text compared to the argumentation rules that are observed in formal discussions, such as political debates and legal affairs. Arguments in informal discourse rarely follow the logical structure of an argument having claims supported by facts, warrants, and qualifiers. On the contrary, they are often implicit without a solid logical structure, thus more agile approaches have to be followed for detecting argumentation in a statement and further analyse it. For example, an argument expressed through a tweet is typically an one-sentence argument expressing a stance supported by a an external resource without requiring any fact-checking.

Having in mind those special characteristics of argumentation in short text, and especially in social media, a new agile definition for argument must be provided covering the needs of this emerging area. In this paper, before elaborating on the essence of argumentation from a computational approach (section 4), a shorter and simpler mathematical expression is provided forming a definition strict enough to exclude any ambiguity, but at the same time fairly agile to be applied in noisy text.

definition

Definition 1. ***Existence of argument*** *is expressed as a quadruple in:* $\langle s_{ijt}, c_{it}, r_{jt}, t\rangle$ *where* s_{ijt} *is the initial statement or set of statements that is supported by* c_{it} *claims (or premises) that are used to support or oppose an idea (or suggestion) for the topic* t *that is questionable or open to doubt using a rationale* r_{jt}*.*

The definition 1 covers a wide range of argumentation's variations in social media discourse since: 1) it narrows down the area of interest to a selected topic t, 2) it identifies the stance (positive, negative) of the claimant through the claim c, and 3) it proceeds -to some extent- with the examination of some urgent topics in the NLP area in social media text such as reason acceptability, facts recognition and rumour detection through the inclusion of reasoning r in the definition. The subscripts ijt depict the direct relations between the components. For example, the claim c is a structural component of the statement s (declared through i) for a specific topic t, but there is not an explicit relation with a reason r.

Considering the noisy nature of text in social media and the challenges it poses, argumentation modelling schemes should seek for novel approaches to the problem to map qualitative characteristics to quantitative features. The proposed modelling schemes should be more agile providing the necessary flexibility to researchers to implement different versions and compare their performances.

4 Argument Quantification

A statement is considered argumentative when it provides reasons for or against the discussed topic. In the context of argumentation in social media, topics are expressed with the use of hashtags and the statements are the tweets. After studying a series of tweets on multiple topics, we concluded that two different rationales fall under this definition: 1) (try to) persuade towards a specific stance, and 2) provide evidence (news media, blog, expert opinion) that supports a stance towards the discussed topic. An example of the given definition is illustrated in Figure 1. The tweet is a statement for the construction of the gas pipeline under the name *#Nordstream2* (topic), expressing a negative stance through a claim (*EU should block*), which is justified by a reason (*on climate grounds*). The argument, although short, is solid and on point since it is compliant with the definition's requirements for the existence of argument. Adjusting the Definition 1 in the wider context of NLP in social media, a tweet is equivalent to a statement, the presence of hashtags poses the limits for the topic selection, the identification of claims is accomplished through the task of stance detection, and the reasoning detection is a combination of tasks such as source identification, rumour diffusion, and reliability evaluation.

🟩 Statement 🟥 Claim 🟦 Rationale

[2] EU should block #NordStream2red!60green!60[2] whitegreen!60[2]on climate grounds blue!60green!60

Fig. 1: An example illustrating the existence of argumentation in short text as given in the Definition 1 .

The reliability of the reasoning is not examined, because the integrity and the soundness of the argument are not in the scope of this work. In the environment of social media, where different topics are discussed, trends appear, and hashtags are created in the blink of an eye, it is important to narrow down the scope of our research by setting limits. On this constantly changing environment, the limits are defined from the topic each statement addresses thus it is important to define the concept of topic.

definition

Definition 2. ***Topic*** *A topic t determines the context of the discourse under examination and defines the dialectic limits that can be applied in a logical acceptable statement. The finite potential topics formulate the set T, and therefore $t \in T$.*

The limits of each t can be differentiated depending on the needs of each case study. A prerequisite for argumentation is the existence of claim(s) in the block of text under examination.

definition

Definition 3. ***Claim*** *A claim c is an assertion containing or implying a stance towards a topic t. The finite potential claims formulate the set C, and therefore $c \in C$.*

Through the presence of a claim, the stance of the claimant towards the topic t can be extracted. The stance has already been defined [19] and although similar to the Definition 1, it is the product of an argument analysis and not an integral part of it.

The goal of the claimants is to persuade their audience, supporting their original claim(s) through a reasoning process that completes the argument. Even in very short text, the existence of justification is still very important, even if it is incomplete or weak.

definition

Definition 4. ***Reason*** *A reason r is a justification that supports or tries to support a specific claim c towards a topic t. The finite potential reasons formulates the set R, and therefore $r \in R$.*

The co-existence of a concrete claim followed by a reason does not necessarily create a legitimate argument, because the provided reason may not support sufficiently the original claim. Therefore, a mapping function assessing the appropriate match of the two objects needs to be defined, demonstrating that every reason r can sufficiently support a specific finite number of claims c.

definition

Definition 5. ***Mapping reason to claim*** *A reason is valid when there is a function $f_{rsn}(r_i, t) : R \rightarrow C$ valuing the validity of a generic $r \in R$ to support a $c \in C$. Then, the valid reasons for a claim $c \in C$ are expressed as a set $Rc = (r_1, t), ..., (r_n, t)$ (for $n \geq 0$) and a mapping function v on Rc is defined as:*

$v(r, c) = g((r_1, t), ..., (r_n, t))$,

where $g : R \times T \rightarrow C$ is a function aggregating and assessing the impact of the existing distinctive valid reasons capable to support assertive claims.

Following the definitions of an argument's components, we proceed on defining the argument itself with the use of computational logic. Assembling the distinctive components of an argument, we provide an argument's definition capable to capture the presence of argumentative discourse in a statement and to be applied on noisy text.

Theorem 1. *An argument a is a statement that supports or tries to support a specific claim c where $c \in C$ towards a topic t, supported by a reason or set of reasons $R_{ct} \in R$ with $R_{ct} \geq 1$ and it is expressed as a set of tuples $\{$(c, r_j), (c, r_{j+l}), ..., (c, r_{m-l}) $\}$ where $0 \leq l < m - j$, $\forall j$ where $1 \leq j < m$ iff $\exists$ c $\in C$ $\bigwedge$ $\exists$ r $\in R_{ct}$: (c, r) $\in A$, where A is a finite set of the potential arguments.*

The above theorem introduces the prerequisites for the creation of an argument adopting a computational approach, but it does not evaluate its validity which is assessed through a function that maps the suitability of the claims to the argument as an entity.

definition

Definition 6. ***Mapping claim to argument*** *A claim is asserted when there is a function $f_{clm}(c_i, t) : C \rightarrow A$ valuing the applicability of a generic $c \in C$ to* *create an $a \in A$. Then, the asserted claims for an argument $a \in A$ are expressed as a set $Ca = (c_1, t), ..., (c_l, t)$ (for $l \geq 0$) and a mapping function u on Ca is defined as:*

$$u(c, a) = h((c_1, t), ..., (c_j, t)),$$

where $h : c \times T \rightarrow A$ is a function aggregating and assessing the impact of the existing statements. The statements are considered arguments because they consist of valid reasons and assertive claims.

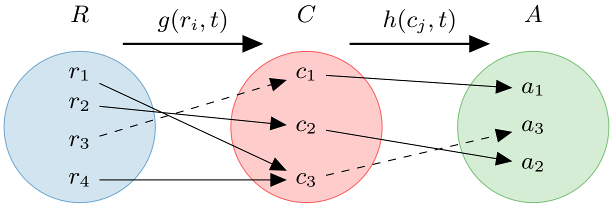

Fig. 2: A graphical illustration of mapping reasons to claims and eventually to arguments

Figure 2 illustrates the mapping process from a set of reasons to a set of claims through a function $g(r, t)$ and from a set of claims to a set of arguments through a function $h(c, t)$. The two functions have as common the input parameters for the same topic t which defines the limits of the sets and both $g(r, t)$ and $h(c, t)$ present $n : 1$ and $1 : 1$ relation from input parameters to output, respectively. The solid lines between the sets indicate a direct link between them, whereas the dotted lines represent an implicit correlation without any evident connection. Neither the argument definition nor the mapping functions do they examine the validity of the argument, but they are limited to its detection and identification of their distinctive components. Therefore, a definition determining the validity of an argument is also required.

Lemma 1. *An argument for a topic $t \in T$ is a tuple of the form $\{c, R_{ct}\}$ where $c \in C$ and $R_{ct} \in R$ is a set of reasons supporting the claim c while $R_{ct} \geq 1$. An argument as a whole is valid if at least one element of the R_{ct} is valid and implies a claim $c \in C$, and c creates an argument a, and it is expressed as $a : (r, r \rightarrow c, c \rightarrow a, a)$.*

Proof. The expression is proved from the law of detachment. For the validity of the claim we have $(r \rightarrow c, r, c)$, and similarly the validity of the argument is proven $(c \rightarrow a, c, a)$.

In other words, an argument is valid when it consists of reason that are based on sensible reasoning and assertive claims. Apart from validity, another important aspect of argumentation is the soundness of an argument; an argument is sound when it is valid, and all of its premises are accurate and truthful. Symbolic logic can be used to check the validity of an argument, but it cannot be used to examine its soundness. The detection of a sound argument is in the scope of research of evidence, including tasks such as source identification, facts recognition and evidence classification.

Finally, having recognized the distinctive components of an argument and provided definitions for each one of them, we introduce the Abstract Framework for Argument Detection (AFAD), a conceptual model that can be applied on (very) short text and detect the presence of argumentative text. The AFAD is the first framework -to the best of our knowledge- that focuses on the detection of argument and not its evaluation, while its structure allows its utilization on short and noisy text. In contrast to previously argumentation frameworks, AFAD offers a great level of flexibility because of its ability to be adjusted on the needs of every experiment, allowing researchers to define the mapping functions.

definition

Definition 7. ***Abstract Framework for Argument Detection (AFAD)*** *is a 4-tuple* $\langle C, R, T, V \rangle$*, where* C *is a finite set of claims that are supported from a finite set of reasons (justifications)* R*. The parameter* T *indicates the topic for which the* C *and the* R *are expressed; a necessary parameter in order to narrow down the potential tuples of* $\langle\ C,\ R\ \rangle$*, and* $V : A \rightarrow C \times R \times T$ *is a function mapping valid arguments to their building components.*

The absence of transition words and the overlapping between claims and conclusions in short text create an environment where original ideas and argumentation schemes can be applied. Therefore, if we wished to juxtapose novel approaches to established argumentation schemes (e.g. Toulmin's model), we should add fillers to reconstruct a complete argument.

5 Methodology

This section presents the methodology that has been followed in this research. More specifically, subsection 5.1 illustrates the annotation process of the dataset, subsection 5.2 presents tehcnical details on the implementation of the rule-based methodology, subsection 5.3 provides details on the implementation of the ML algorithms, and, finally, subsection 5.4 introduces a hybrid architecture for the task of argumentation detection.

5.1 Annotation scheme and corpus creation

Taking into consideration the growing importance of the social media's role in different aspects of everyday life, including social and political debates, harvesting social media could offer a valuable source of data. In this project we are focusing on the geopolitical debate on Twitter about the expansion of the "Nord Stream" gas pipeline in northeast Europe, under the name "Nord Stream 2". The construction of the pipeline has primarily financial incentives, but it also raises concerns on political, environmental, and ethical issues, as there is a conflict of interest between parties and bodies that act for both national and European interests [14]. Concepts such as energy union and energy diversification are manipulated depending on the goal of each party and are used to create strong arguments for or against the construction of the pipeline.

Example	Annotation
Read our in-depth weekly overview on European natural gas matters http:// buff.ly/1N5bBr0 #StateEnUn15 #Groningen #NordStream2	Non argumentative
Report: German Finance Minister teams up with Putin on #NordStream2 http:// buff.ly/1MQzItm	Non-argumentative
Deeply concerned about Germany's lack of concern with EU law for #NordStream2. RT if you are deeply concerned, too. @AuswaertigesAmt	Argumentative
What would be your question for high German official about #NordStream2, #energy policy #energyunion ? Thanks Twitter. :-)	Non-argumentative
U.S. LNG exports to EU, a powerful potential alternative to Russian-sourced energy. But is EU married to Gazprom? #GIPL #NordStream2	Argumentative

Table 1: Examples of Argumentative or Non-Argumentative

Table 1 presents five examples of the collected tweets alongside with their annotations offering an insight into the available information. The examples fall into different categories (news title, ironic tweet, open question) presenting examples from different perspectives. The first two examples are news titles directing to an external resource, the third one is a personal statement expressing disappointment for a specific action, the fourth one opens a thread to collect questions on the topic, and the last statement expresses a negative stance on the debate topic using irony. The number of tweets (∼45000) that have been written and the number of accounts (∼7000) that have been expressed concerning the construction of the Nordstream2 pipeline reveal the great interest of the audience for this trasnational debate. Two of the authors participated in the annotation process and a dataset with 590 annotated statements was created having 452 statements annotated as non-argumentative and 138 of them as argumentative. It is a highly skewed dataset which has opposite skewness compared to the one observed in [12], where the 77.3% of the collected tweets express an argumentative stance. The inter-annotator observed agreement reached 87.2%, and the unweighted Cohen's Kappa score is 0.64, which is considered a good degree of agreement [17].

Table 2 presents the IAA of related work which varies in the interval [0.6, 0.8] indicating substantial agreement. However, related tasks in the wider area of argumentation mining can present higher variance depending on both the task that takes place and the nature of the dataset. For example, for the task of matching argument to specific key points [4], a moderate agreement (0.5 Cohen's kappa) is achieved, whereas for the tasks of reasoning revision on argumentative essays [3] and identification of argumentative components in medical abstracts [20] moderate agreement is achieved, with 0.75 Cohen's kappa and 0.72 Fleiss' kappa respectively.

5.2 Rule-based Approach

The concept of Argument Base (AB) [9] includes a collection of arguments and relations, where each argument formalises knowledge conducive to solve the problem in question. Combining this approach with the provided definition of argument in short text, we separate AB into two entities, claim database (claim DB) and reason database (reason DB), and disregard potential relations while

Authors	Topic	Score
Lytos et al.	Nord Stream 2	0.64 Ck
Addawood et al. [1]	Apple/FBI encryption debate	0.67 Ck
Bosc et al. [6]	5 debate topics	0.81 Ka
Dusmanu et a. [12]	Grexit / Brexit	0.77 Ck

Table 2: A comparison of IAA of related work for the task of argumentation detection.

focusing on the detection of argumentative text. The claim DB includes common claims that are expressed in the debate and the reason DB contains the reasoning that supports the claims.

The rule-based approach introduces a method capable of identifying the conceptual proximity between previously unseen chunks of text and the collection of claims/reasons. For example, the tweet *EU should block #NordStream2 on climate grounds* and the claim *the gas pollution is the main threat for the biodiversity of the Baltic sea* are correlated, even though they do not share any common words. For the construction of the claim DB and the rule DB, claims and reasons are manually extracted after the evaluation process during the annotation task and they express the main objectives of the public debate. Consequently, the knowledge stored in the AB is domain-dependent and it cannot easily be transferred to new domains.

The correlation among a tweet and the collected claims/reasons (mapping functions) can be estimated in multiple ways. However, a sequence of symbols, has limited usefulness as they cannot be fed directly to many algorithms because most of them expect numerical feature vectors with a fixed size and not a series of characters with variable length. For this reason, a processing pipeline has been designed. The very first step of the pipeline is a pre-process function that eliminates the evident noise such as stopwords, URLs, leading and trailing whitespaces. On the other hand, hashtags and mentions are striped (e.g. #NordStream became NordStream), because hashtags can contain claims or reasons (e.g. #climatechange) while mentions can be used as citations to authority. The tokens that have been collected construct the dictionary which is expanded whenever a new entry is introduced in the dataset. Eventually, a corpus of phrases/sentences is represented by a matrix having as rows the tweets of the dataset and as columns the tokens of the dictionary.

The process of turning a collection of phrases/sentences into numerical feature vectors is called vectorization, and there are different ways for expressing the relation of a phrase/sentence with a concept. The most straightforward solution, term frequency (TF), counts the recurrences of each token while completely ignoring the relative position information of the words in the corpus. This method could lead to overtraining due to the repeatability of specific terms in a specific domain without any significance. A technique for weighting different terms in the vectorization process is the term frequency, inverse document frequency (TF-IDF) [15]. The first part is the number of occurrences of the term in the document, and the second part is the inverse of the number of documents that contain the specific term. Therefore, the specificity of a term can be quantified as an inverse function of the number of documents in which it occurs. For this research study, both approaches have been followed and a series of comparisons are presented.

However, the challenge lies not not only in the transformation of the text into a binary format that a machine can understand, but also in measuring the semantic similarity of the constructed vectors. A way of estimating the conceptual proximity between a tweet in the dataset and an argument in the AB is the calculation of the cosine similarity between two vectors. Let $\mathbf{t}$ and $\mathbf{a}$ be two vectors of the same size representing a tweet in the dataset and an argument in the AB, respectively. Using the cosine measure as a similarity function, we have:

$$sim(\mathbf{t}, \mathbf{a}) = \frac{\mathbf{t} \cdot \mathbf{a}}{||\mathbf{t}|| ||\mathbf{a}||} \tag{1}$$

where $\cdot$ is the dot product of the two vectors, and $||\mathbf{t}||$ is the Euclidean norm of vector $\mathbf{t} = (t_1, t_2, ..., t_n)$, defined as $\sqrt{(t_1^2 + t_2^2 + \ldots + t_n^2)}$. Similarly, $||\mathbf{a}||$ is the Euclidean norm of vector $\mathbf{a}$. The multiplication of the two vectors is achieved using the transpose operator allowing us to multiply the components of the two vectors by flipping the transposed vector:

$$\mathbf{ta^T} = \begin{bmatrix} t_1, t_2, ..., t_4 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_4 \end{bmatrix} = t_1 a_1 + t_2 a_2 + ... + t_n a_n \tag{2}$$

The measure computes the cosine of the angle between vectors t and a, returning values between [0, 1]. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal), thus they do not have any semantic correlation. On the other hand, the closer the cosine value is to 1, the smaller the angle and the greater the correlation between the two vectors. Negative values do not exist because the encoding methods we have selected do not assign negative values to the text.

5.3 ML Approach

ML classification algorithms automatically classify a set of previously ”unseen” text segments to a set of predefined class labels based on previously labelled data on which the algorithms have been trained. This process requires resources that are related to the given task as well as useful features that can be extracted from either plain text or metadata. Social media are typical resources that can provide a large amount of user-generated data with semi-structured or structured metadata. The same pre-process function that has been deployed in the rule-based component has also been applied here. A wide range of statistical and linguistic features have been suggested for argumentation detection and other NLP tasks, such as sentiment analysis and source identification. In this research, a wide set of features has been chosen covering the following categories:

Lexical Features refer to the n-gram encoding technique (uni-gram, bi-gram, etc.). Uni-gram and bi-gram encoding techniques have been deployed with both TF and TF-IDF techniques, creating four comparable clusters.

Semantic Features define the characteristics of the language that can provide a deeper insight into the data such as Part-of-Speech (PoS), dependency relations syntactic and parse trees. The NLTK PoS tagger [5] has been used which can identify and group words into different categories that display similar syntactic behaviour.

Sentiment features are those that reveal emotions and they are usually detected with the use of external lexicons. The textblob software [24] has been used for the extraction of two features: polarity and subjectivity. The former returns a float within the range [-1.0, 1.0], and the latter within the range [0.0, 1.0]. The degree of polarity can be interpreted as negative/positive stance towards a specific topic, whereas high subjectivity score correlates with opinionated claims.

Twitter-specific features are offered as metadata through the Twitter API and concern the specific characteristics a tweet has, such as the length of a message, the presence or not of URLs, mentions of other users, hashtags, and official account verification. Based on these characteristics, binary variables have been used indicating the existence of mentions and hashtags, as well as a counter variable for the characteristics.

Algorithm 1 Execution of ML Algorithms

```
A list of tuples; Define $n$ for text n-gram encoding  Define list of external features
Define list of algorithms  Pre-process the dataset
**while** i $\leq$ n **do** Encode dataset with n-gram  algorithm in algorithm list combination
in feature list Execute *algorithm* with *combination* in *dataset*  Cross-validate class
prediction  Store the results
```

Using different combinations of the aforementioned features, four ML classifiers have been trained for the task of argument detection. Algorithm 1 illustrates the execution of the designed pipeline. For the execution of the ML algorithms, both different text encoding approaches (defined by the variable n) and different combination of features have been tested. The list of features has been limited to the use or not of all the external features, after a preliminary analysis. Eventually, the iterative execution of the ML algorithms returns a list of results based on the different combinations that have been performed. Due to the limited size of the dataset, cross validation has been used preventing over-fitting errors and providing higher reliability in the results.

The algorithms which have been selected to be used in this research represent a different algorithmic approach. In total, four different ML algorithms were deployed and executed with eight different combinations of features providing a wide test area for their performance in the specific task. The chosen algorithms with a short description are below:

Multi-layer Perceptron (MLP) belongs in the family of the Artificial Neural Networks (ANN) and it is based on a function $f(\cdot) : R^m \rightarrow R^o$ that is trained on a given dataset, where m is the number of dimensions (features) for input and o is the number of dimensions for output (two alternatives in our case-study). The transformation of the input to the desired outcome is realised through an activation function. In this research work, the Rectified Linear Unit (ReLU) has been deployed in the hidden layers and the logistic sigmoid function in the output layer. The ReLu has been selected due to its non-saturation capabilities and its low computational requirements and the logistic due to the binary nature of the problem. The two main disadvantages of the ANN algorithms is the unexplained behaviour of the network and the complexity level for the tuning of its parameters; their values with a short description is illustrated in Table 3.

Parameter	**Value**	**Short description**
hidden layer sizes	(100,1)	Number of hidden layers(1) and units(100)
activation function	ReLU	Piecewise linear function defining the output of each unit
output function	Logistic	Logistic sigmoid function defining the output of the last layer
solver	Adam	Stochastic gradient optimization algorithm
L2 penalty	0.0001	Weight penalty set to Adam solver to avoid overfitting
batch size	200	Size of batches used in Adam solver
learning rate	constant	Fixed learning rate
learning rate init	0.001	The step-size for weights update
max_iter	200	Number of epochs for Adam solver
shuffle	True	Shuffle samples in each iteration
random state	0	Define random number generator for reproducible results
tolerance	1e-4	Minimum acceptable change in loss or score
n_iter_n_change	10	Maximum number of epochs to not meet tolerance improvement
beta_1	0.9	Exponential decay rate for the 1st moment
beta_2	0.999	Exponential decay rate for the 2nd moment
epsilon	1e-8	Stabilizer value to prevent any division by zero

Table 3: The parameters that have been used for the deployment of the MLP.

Decision Tree (DT) is a non-parametric supervised learning method which predicts the values of the required feature through a series of decisions rules inferred from the dataset's features. Due to the design of the decision rules, it is capable to handle features that indicate categorical data such as the existence or not of specific hashtags or mentions. Another advantage of the DT solutions is the easy interpretation of the algorithm's flow. On the negative side, DT tends to present a skewness towards the majority class, stuck in local optima and overfit on a dataset with many features. Therefore, its performance on imbalanced data for complicated problems could be questionable.

Logistic Regression (LR) is a statistical classification algorithm that is used to model the class' probability of the required feature and then convert the probability to log-odds through the logistic function. The logistic function that was used for this research is the Limited-memory BFGS (L-BFGS) from the family of quasi-Newton methods [7]. LR is highly interpretable and computationally inexpensive, however, it cannot solve non-linear problems. Therefore, its application on noisy environment with substantial noise could be questionable.

Support Vector Machines (SVMs) are a set of supervised learning methods that represent the samples as points in space and through a mapping process, the points are assigned into classes, producing a non-probabilistic binary linear classifier. Moreover, SVMs can also perform non-linear classification using different kernel functions which map the input into high-dimensional feature spaces. However, the number of features that have been extracted for this research work is large, hence there is no need to map data to a higher dimensional space and the linear kernel has been used. Another characteristic of the SVMs algorithms is that they need a clear margin of separation between classes to outperform other solutions; a prerequisite that is tough to be met on complex or vague classification tasks.

5.4 A combination of approaches

The objective of our hybrid methodology is the modification of a supervised ML pipeline capable of fixing possible bias of the algorithms towards the majority class. The combination of the rule and ML-based approaches have the potential to create a hybrid solution capable of integrating the positive aspects of each one. Under this rationale, the designed pipeline forwards the predictions of the ML algorithms to the rule-based component which is responsible to fix any anomalies that are identified.

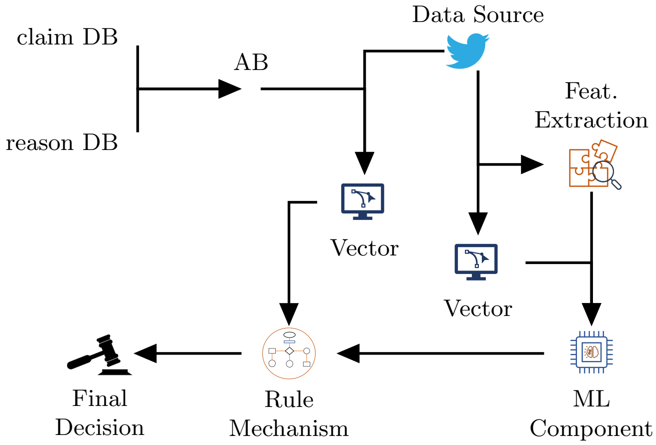

Fig. 3: A graphical illustration of the hybrid approach that has been followed.

The proposed architecture is depicted in Figure 3 illustrating the combination of the ML and rule-based approaches. The ML component receives as input the extracted features (lexical, semantic, sentiment, Twitter-specific) from Twitter, which is used as the main data source. Afterwards, the ML algorithms process the collected data trying to find patterns and eventually determine the class for each sample. For the rule-based approach, the claim DB and the reason DB are normalized into a common format, and alongside with Twitter data are fed into the vectorization algorithm. The predictions of the ML components are compared to solutions suggested from the rule-based mechanism and when there is a controversy between the two predictions, the decision mechanism is enabled. The decision mechanism is based on the idea that the ML algorithms tend to underestimate the minority class while the rule-based mechanism is built to identify even implicit arguments since they present a correlation with stored information in the AB. In terms of performance, it is translated to high precision for ML algorithms and high recall for the rule-based algorithms. For highly negatively imbalanced datasets, such as the Nord Stream 2 dataset, it is important to increase the correct prediction of the positive instance, even if it means decreased performance in the majority class. Therefore, the hybrid decision mechanism keeps the positive ML predictions while in other cases, the rule-based suggestions are preferred due to the domain knowledge that carry. Algorithm 2 illustrates the hybrid solution utilizing the concept for both semantic similarity and ML.

Algorithm 2 Execution of hybrid solution

A list of tuples Receive tuple *texts/predictions* from Algorithm 1 Receive AB Vectorize *text* and AB text-prediction in tuple texts/predictions claim in claim DB reason in reason DB Find semantic similarity between *text* and *claim* Find semantic similarity between *text* and *reason* Estimate total *semantic similarity* Correlate *total semantic similarity* with *prediction* Draw final decision

Finally, the sensitivity of the rule-based mechanism should be noted as it can be easily tuned and thus render any potential adjustments under control. For this research study, the rules have been designed to identify the instances in the minority class and therefore increased recall is expected.

5.5 Evaluation Process

The evaluation for the algorithm's performance is achieved through the use of three different metrics (Precision, Recall, F1-score), each one presenting a different aspect of the algorithm's behaviour. Furthermore, the F1-score has been calculated using different techniques, aiming at covering any doubts that may arise due to the imbalanced nature of the data. Precision is the ability of the classifier not to label as positive a sample that is negative. When the priority is the detection of positive instances on a negatively imbalanced dataset, the importance of precision declines. Recall, on the other side, is the ability of the classifier to find all the positive samples. In many real-life scenarios when the positive class is the minority one, the goal is to increase the recall because it discovers more positive instances, which often, is the objective. Finally, the weighted average of the precision and recall is the F1-score and it is considered the most reliable method. However, on imbalanced datasets reporting exclusively the F1-score is often not enough and it could be misleading. Therefore, three different approaches for its calculation have been implemented:

- Binary: reports the results for the positive class (argumentative), ignoring the performance on the negative one (non-argumentative). It could be misleading when applied on imbalanced dataset due to dominance of the majority class.
- Micro: calculates performance globally by counting the total true positives, false negatives, and false positives. It takes into consideration label imbalance. However, it favours the performance of the algorithms on the more

populated class since it assigns weight on each class based on the number of instances.
- Macro: calculates performance for each label and finds their unweighted mean without taking into consideration the label imbalance. When the minority class is valued the most, it is usually the preferred way of calculation because it treats both classes as equal regardless the number of instances of each class.

6 Experiments and Results

This section presents the results that have been produced from the three different approaches that have been implemented in this research for the task of argumentation detection. More specifically, subsection 6.1 presents the performance of the rule-based approach, subsection 6.2 presents the performance of the ML algorithms, and, finally, subsection 6.3 demonstrates the results of the proposed hybrid architecture.

6.1 Rule-based approach

For the first set of experiments, the rule-based approach as described in subsection 5.2 was implemented, presenting different implementations for the functions $g(r_i, t)$ and $h(c_j, t)$ that represent the mapping from reasons to claims and from claims to arguments, respectively. The first scoring function, $sem\_AB(x)$ function, consists of two distinct tasks; the first computes the semantic proximity between the text and the set of reasons and the second one computes the semantic proximity between the text and a set of claims. Both sets have been created manually from the authors and express -to some extent- the spirit of the debate. The $sem\_AB(x)\_idf$ function follows the same principles but the encoding of the text takes place using the TF-IDF method. The third alternative uses an external general-purpose dictionary [24] that assigns values within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. After the annotation process, it has been observed that tweets without any subjectivity aspects are usually news titles that redirect to external resources hence higher subjecitivity is correlated with higher probability on expressing arguments. Finally, for the last approach, a combination of the two methods was implemented ($comb(x)$ function), where a chunk of text has to display both semantic similarity with a known argument and reach a relatively high subjectivity score. Apart from the different functions, two more methods were also used as benchmarks. The first benchmark is a random function, and the second one is the Jaccard similarity coefficient score (or Jaccard index). Jaccard index is used for gauging the similarity and diversity of sample sets, and it is defined as the intersection between the arguments stored in the AB and the collected tweets divided by the size of the union of the sample sets: $J(AB_i, T_j) = \frac{|AB_i \cap T_j|}{|AB_i \cup T_j|}$.

Table 4 presents the results produced after the deployment of the six different rule-based techniques. The first column has the name of the deployed technique

	Prec	Rec	F1 score	F1 (micro)	F1 (macro)
Baseline	0.25	0.56	0.35	0.53	0.49
Jaccard Sim.	0.28	0.64	0.39	0.56	0.52
sem_AB(x)	0.34	0.77	0.47	0.62	0.59
sem_AB(x)_idf	0.33	0.73	0.45	0.61	0.57
sub(x)	0.27	0.61	0.38	0.55	0.51
comb(x)	0.37	0.45	0.41	0.70	0.61

Table 4: Comparison of the different rule-based mechanisms that have been applied. The metrics that have been used are precision, recall, and f1-score calculated with three different estimation methods.

and the other five columns have precision, recall and F1-score calculated with three different methods. For the identification of the argumentative tweets, the $sem\_AB(x)$ function surpasses the rest of the methods with 0.47 f1-score, followed by the $sem\_AB(x)\_idf$ with 0.45. The general-purpose lexicon comes up short on every calculation technique, but it is better than the random baseline, with 0.38 compared to 0.35 f1-score. The $comb(x)$ function fails to incorporate the benefits out of them and presents the highest precision, but also a great drop in the recall. Finally, the Jaccard index comes up short with 0.39 f1-score, suggesting that argumentation detection is a complicated task that requires more sophisticated approaches, however, it outperforms the random baseline indicating its suitability as an advanced baseline.

6.2 ML-based approach

The second set of experiments includes the execution of four different ML algorithms with different encoding mechanisms (TF and TF-IDF) on both uni-gram and bi-gram level. Moreover, the algorithms are executed with and without additional features (semantic, sentiment, twitter-specific) assessing the capability of the algorithms to exploit these extra features. The results provide a good overview for the performance of different ML algorithms for the task of argumentation detection.

Table 5 presents the F1-score calculated with binary, micro, and macro calculation for four different algorithms. The best performance is achieved when the MLP is deployed using TF uni-gram encoding with external features reaching 0.50 f1-binary and 0.69 f1-macro score. The second best performance is observed when the SVM is executed with TF uni-gram encoding while using external features, presenting 0.47 f1-binary and 0.67 f1-macro score. The impact of the encoding method seems to be the most important factor since every algorithm presents -almost- always better results when they receive text encoded with the TF technique. Additionally, the use of bi-gram does not seem to offer any additional value in the classification task. The limited length of the text and the protection of the special characteristics of Twitter (e.g. hashtags, mentions) sets the use of TF-IDF encoding ineffective. Regarding the use of additional features, even though the highest and the second-highest score is achieved using additional

			F1-binary	F1-micro	F1-macro	F1-binary	F1-micro	F1-macro
N-gram				Encoding			Encoding + Features	
8*Uni-gram	4*TF	MLP	0.45	0.81	0.67	0.50	0.80	0.69
		DT	0.35	0.76	0.60	0.28	0.75	0.56
		LR	0.38	0.80	0.63	0.40	0.81	0.64
		SVM	0.44	0.79	0.65	0.47	0.79	0.67
	4*TFIDF	MLP	0.46	0.80	0.67	0.46	0.79	0.67
		DT	0.31	0.75	0.58	0.27	0.73	0.55
		LR	0.00	0.77	0.44	0.20	0.79	0.54
		SVM	0.25	0.80	0.57	0.29	0.79	0.58
8*Bi-gram	4*TF	MLP	0.41	0.81	0.65	0.26	0.80	0.58
		DT	0.30	0.75	0.58	0.26	0.75	0.55
		LR	0.28	0.80	0.58	0.34	0.81	0.61
		SVM	0.43	0.82	0.66	0.42	0.81	0.65
	4*TFIDF	MLP	0.43	0.80	0.65	0.41	0.81	0.65
		DT	0.35	0.76	0.60	0.27	0.74	0.56
		LR	0.00	0.78	0.44	0.13	0.78	0.50
		SVM	0.15	0.79	0.51	0.25	0.79	0.56

Table 5: Comparison of the ML algorithms' performance when applied on different encoding techniques and external features are used. The suggested values in the algorithms' parameters as provided from the sklearn [22] are used. Exception is the use of linear kernel for the SVC.

features, no clear conclusion can be drawn. It seems that the use of excessive complicated encoding does not provide the expected results.

6.3 Hybrid approach

The last set of experiments includes the execution of the hybrid solution which is the combination of ML algorithms followed by revision from the rule-based component. The performance of the ML algorithms is impressive when estimated with micro calculation, thus the rule-based component aims at increasing the F1-score in binary calculation. Similarly to the ML-based approach, the algorithms are executed with different encoding mechanisms and there are execution batches that include the additional features. The hybrid solution is expected to have higher recall compared to ML-based solutions due to the addition of positive instances which are recognized from the rule-based component while the precision of the algorithms is expected to decrease.

Table 6 presents the f1-score with binary, micro, and macro calculation for the hybrid solution. In terms of f1-binary score, every hybrid solution outperforms its corresponding plain ML implementation. The performance of the proposed hybrid solution surpasses the ML solution, independently from the algorithms that have been deployed while offering a minimum standard for every deployed algorithm which is at least equal or surpass the performance of the ML solution when estimated with binary calculation. The best performance is achieved when the MLP is deployed using TF-IDF uni-gram encoding reaching 0.54 f1-binary

			F1-binary	F1-micro	F1-macro	F1-binary	F1-micro	F1-macro
N-gram			Hybrid			Hybrid+		
8*Uni-gram	4*TF	MLP	0.52	0.73	0.67	0.53	0.72	0.67
		DT	0.48	0.69	0.63	0.46	0.69	0.62
		LR	0.47	0.71	0.64	0.48	0.72	0.64
		SVM	0.50	0.71	0.65	0.51	0.71	0.65
	4*TFIDF	MLP	0.54	0.73	0.68	0.53	0.72	0.67
		DT	0.46	0.69	0.62	0.44	0.67	0.60
		LR	0.40	0.70	0.60	0.42	0.70	0.61
		SVM	0.46	0.72	0.63	0.46	0.71	0.63
8*Bi-gram	4*TF	MLP	0.52	0.74	0.67	0.48	0.73	0.65
		DT	0.48	0.69	0.63	0.47	0.70	0.63
		LR	0.46	0.72	0.63	0.47	0.72	0.64
		SVM	0.51	0.73	0.66	0.51	0.73	0.66
	4*TFIDF	MLP	0.52	0.73	0.67	0.50	0.73	0.66
		DT	0.45	0.68	0.61	0.44	0.68	0.61
		LR	0.41	0.70	0.61	0.40	0.70	0.60
		SVM	0.44	0.71	0.63	0.45	0.71	0.63

Table 6: Comparison of the hybrid solutions' performance when applied on different encoding techniques and external features are used.

score. On the other hand, when the performance is estimated using micro calculation, the ML algorithms outperform the hybrid solutions and when the macro average is used, the two approaches present comparable results. Micro-averaged results are a measure of effectiveness for the majority class, which in our case study is the less important class. Overall, there is marginal deviation between the alternatives that have been deployed, around 0.48, 0.70 and 0.60 in binary, micro and macro calculation respectively.

The results present an impressive consistency indicating the significant impact of the rule-based mechanism on the hybrid solution while the impact of external features into the ML algorithms was reduced to a minimum level for the hybrid solution. The impact of the hybrid approach is evident when different encoding methods are compared since neither the encoding technique nor the number of tokens that are included in the encoding process have a significant impact on the algorithm's performance.

Moreover, the results of the algorithms indicate the impact of the hybrid solution which increases the performance of -almost- every solution, but it eliminates the unique behaviour of each algorithm, hence it is important to delve deeper into the behaviour of the hybrid solution compared to the ML algorithms. Table 7 presents the difference in the performance of the two algorithms with the hybrid solution outperforming the ML one in terms of recall, but come short in precision. For example, the MLP executed with unigram TF without any features achieves 0.61 precision and 0.36, whereas when it is integrated into the hybrid architecture it presents 0.43 precision and 0.67 recall.

			Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
			ML solution				Hybrid solution			
N-gram			Encoding		Encoding + features		Encoding		Encoding + features	
8*Unigram	4*TF	MLP	0.61	0.36	0.57	0.45	0.43	0.67	0.43	0.70
		DT	0.45	0.28	0.38	0.22	0.38	0.62	0.38	0.59
		LR	0.61	0.27	0.66	0.29	0.40	0.58	0.41	0.58
		SVM	0.53	0.37	0.53	0.42	0.41	0.64	0.40	0.67
	4*TF-IDF	MLP	0.60	0.38	0.55	0.39	0.44	0.70	0.43	0.70
		DT	0.41	0.25	0.34	0.22	0.37	0.58	0.35	0.58
		LR	0.00	0.00	0.62	0.12	0.36	0.45	0.37	0.48
		SVM	0.71	0.15	0.57	0.20	0.40	0.55	0.40	0.56
8*Bi-gram	4*TF	MLP	0.70	0.30	0.78	0.16	0.44	0.65	0.42	0.58
		DT	0.41	0.24	0.38	0.20	0.38	0.62	0.39	0.59
		LR	0.68	0.17	0.70	0.23	0.40	0.54	0.40	0.56
		SVM	0.74	0.30	0.64	0.32	0.43	0.62	0.42	0.64
	4*TF-IDF	MLP	0.61	0.33	0.69	0.29	0.43	0.67	0.42	0.62
		DT	0.44	0.30	0.37	0.22	0.36	0.58	0.36	0.57
		LR	0.00	0.00	0.50	0.08	0.37	0.45	0.36	0.45
		SVM	0.69	0.08	0.60	0.16	0.39	0.52	0.39	0.54

Table 7: Comparison table between ML-based solution and hybrid solution in terms of precision and recall.

7 Discussion and Challenges

The act of argumentation takes place in our effort to impart our views or analyse and break down the premises of the arguments of others. This explains why argumentation modelling initially focused on rhetorics, academic text, and political debates. In the meantime, the developments beyond Web 2.0 have eliminated the boundaries between the physical and digital world while social media have transformed the means of communication and information exchange. New technical challenges with strong social implications have emerged such as fake news and hate speech detection.

The proposed hybrid solution combines the benefits of the data-driven solutions using ML algorithms while integrating domain knowledge through a rule-based mechanism. In this regard, the proposed method can identify implicit arguments that cannot be detected from the ML algorithms, a very important ability when imbalanced datasets are used due to the tendency of the ML algorithms to favour the dataset's majority class. Furthermore, by formalizing the definition of argumentation in short text, the theoretical foundations for different development strategies are provided. The concept of semantic similarity for the task of argument detection is assessed providing domain knowledge through a limited AB. However, the manual construction of the AB poses two major challenges; the rise of scalability issues because of limited coverage and the risk of bias because the AB is constructed after the examination of the Twitter dataset.

Moreover, a series of experiments are executed through different combinations of text encoding methods and algorithms while also using different evaluation metrics to gain a complete overview of the proposed solution. Two critical concerns were raised by the deployment of the hybrid solution; the drop in the solution's performance when evaluated with micro evaluation and the strong impact of the rule-based mechanisms that obscure the unique characteristics of the ML algorithms. Based on the experimental results, the following insights can be compiled from the experiments:

- Imbalanced data express real-world scenarios and often require a special approach, thus the use of different estimation methods (i.e. binary/micro/macro calculation) for the metrics that are used is of crucial importance for having a holistic view of the problem.
- The results from the execution of the rule-based approach indicate that simple methods such as Jaccard index for estimating the semantic similarity cannot be applied in complex tasks like argument detection.
- For the creation of a balanced rule-based mechanism, attention should also be paid on the negative class because otherwise, the precision of the solution will be reduced more than the expected threshold.
- Vectorization techniques are of major importance for the argument detection task since their effect is significantly stronger compared to the value that is offered from the additional features.
- Hybrid solutions present better results in imbalanced data due to their capability of identifying instances in the minority class, which is typically the most important one in real-world applications.

Despite the challenges that have been raised, the performance of the hybrid solution is more than promising while domain knowledge can reveal knowledge aspects in the minority class. Moreover, the creation of the rule-based mechanism offers a calibration process that can enhance the precision or the recall of the hybrid solution depending on the task at hand, and thus offer tailored solutions based on the nature of the problem.

8 Conclusions and future work

In this paper, we have presented a new approach for detecting argumentation in short text in real-life settings. First, we introduced a formal definition of argumentation providing the necessary theoretical background for different implementation. Second, we created an annotated dataset for the task of argument detection achieving substantial agreement on a previously unexplored field. Third, we proceeded to a comparison of rule-based solutions testing different methods and their impact on different metrics. Moreover, we implemented an extensive comparison among ML algorithms using different encoding mechanisms, showing that the encoding technique is the most important factor surpassing the impact of any additional features that are used. Finally, we proposed a hybrid solution which can increase the recall capability for every ML algorithm that has been deployed regardless of the encoding technique that has been used.

The findings of this research can be used as a point of reference for future studies, either by exploring different implementation methods in the proposed modelling scheme or through the comparison in the performance of alternative solutions. The extensive comparisons between methods provide a good overview for state-of-the-art solutions. Finally, the dataset that has been created provides a perspective for real-life settings where the positive class is the minority one creating additional challenges.

In future work, we will explore the applicability of the proposed methodology on new domains, through the gained knowledge transfer. We are interested in enhancing the manual process of the AB construction with the use of automatic argument discovery approaches that are capable of identifying arguments in previously unseen text. The task of argument discovery presents many similarities to topic modeling, thus unsupervised algorithms will be integrated into the existing argumentation detection pipeline. Finally, another consideration for future work is to explore more sophisticated approaches for semantic similarity.

References

1. Addawood, A.A., Bashir, M.N.: What is Your Evidence? A Study of Controversial Topics on Social Media. In: Proceedings of the 3rd Workshop on Argument Mining. pp. 1–11. Berlin, Germany (2016)
2. Afoudi, Y., Lazaar, M., Al Achhab, M.: Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. Simulation Modelling Practice and Theory **113**, 102375 (12 2021). https://doi.org/10.1016/J.SIMPAT.2021.102375
3. Afrin, T., Wang, E., Litman, D., Matsumura, L.C., Correnti, R.: Annotation and Classification of Evidence and Reasoning Revisions in Argumentative Writing. In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 75–84. Seattle, WA, (2020), `https://www.aclweb.org/anthology/2020.bea-1.7`
4. Bar, R., Lilach, H., Friedman, E.R., Kantor, Y., Lahav, D., Slonim, N.: From Arguments to Key Points: Towards Automatic Argument Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4029–4039. Association for Computational Linguistics (2020), `https://arxiv.org/abs/2005.01619`
5. Bird, S., Edwardm Loper, Ewan, K.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
6. Bosc, T., Cabrio, E., Villata, S.: Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. In: Proceedings of the 6th International Conference on Computational Models of Argument. pp. 21–32. Potsdam, Germany (9 2016)
7. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A Limited Memory Algorithm for Bound Constrained Optimization. SIAM Journal on Scientific Computing **16**(5), 1190–1208 (9 1995). https://doi.org/10.1137/0916069
8. Cabrio, E., Villata, S.: Five Years of Argument Mining: a Data-driven Analysis. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. pp. 5427–5433. International Joint Conferences on Artificial Intelligence Organization, California (7 2018). https://doi.org/10.24963/ijcai.2018/766
9. Carstens, L., Toni, F.: Using Argumentation to Improve Classification in Natural Language Problems. ACM Transactions on Internet Technology **17**(3), 1–23 (7 2017). https://doi.org/10.1145/3017679

10. Castano, S., Falduti, M., Ferrara, A., Montanelli, S.: A Bootstrapping Approach for Semi-Automated Legal Knowledge Extraction and Enrichment. In: SEBD 2020. pp. 1–11. Villasimius, Sardinia, Italy (6 2020)
11. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (9 1995). https://doi.org/10.1016/0004-3702(94)00041-X
12. Dusmanu, M., Cabrio, E., Villata, S.: Argument Mining on Twitter: Arguments, Facts and Sources. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2317–2322. Copenhagen, Denmark (2017)
13. Eger, S., Daxenberger, J., Stab, C., Gurevych, I.: Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need! In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 831–844. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018), http://www.ukp.tu-darmstadt.de
14. Fischer, S.: Lost in regulation: The EU and Nord Stream 2. CSS Policy Perspectives **5**(5) (2017). https://doi.org/10.3929/ethz-b-000210438, https://doi.org/10.3929/ethz-b-000210438
15. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval (1972). https://doi.org/10.1108/eb026526
16. Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N.M., Wu, Y.: Exploring the Limits of Language Modeling. Tech. rep., Google (4 2016)
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–74 (3 1977), http://www.ncbi.nlm.nih.gov/pubmed/843571
18. Lytos, A., Lagkas, T., Sarigiannidis, P., Bontcheva, K.: Argumentation Mining: Exploiting Multiple Sources and Background Knowledge. In: 12th Annual South-East European Doctoral Student Conference (DSC2018). pp. 66–74. Thessaloniki, Greece (5 2018), https://arxiv.org/abs/1809.06943v1
19. Lytos, A., Lagkas, T., Sarigiannidis, P., Bontcheva, K.: The evolution of argumentation mining: From models to social media and emerging tools. Information Processing and Management (2019). https://doi.org/10.1016/j.ipm.2019.102055
20. Mayer, T., Cabrio, E., Villata, S.: Transformer-based Argument Mining for Healthcare Applications. In: ECAI. pp. 2108–2115 (8 2020), https://hal.archives-ouvertes.fr/hal-02879293https://hal.archives-ouvertes.fr/hal-02879293/document
21. Ou, C., Jin, X., Wang, Y., Cheng, X.: Modelling heterogeneous information spreading abilities of social network ties. Simulation Modelling Practice and Theory **75**, 67–76 (6 2017). https://doi.org/10.1016/J.SIMPAT.2017.03.007
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(Oct), 2825–2830 (2011), http://www.jmlr.org/papers/v12/pedregosa11a.html
23. Peldszus, A., Stede, M.: From Argument Diagrams to Argumentation Mining in Texts. International Journal of Cognitive Informatics and Natural Intelligence **7**(1), 1–31 (1 2013). https://doi.org/10.4018/jcini.2013010101
24. Steven Loria: TextBlob: Simplified Text Processing — TextBlob 0.15.1 documentation
25. Sun, S., Luo, C., Chen, J.: A review of natural language processing techniques for opinion mining systems. Information Fusion **36**, 10–25 (7 2017). https://doi.org/10.1016/j.inffus.2016.10.004

26. Toulmin, S.E.: The Uses of Argument. Cambridge University Press, Cambridge (2003). https://doi.org/10.1017/CBO9780511840005
27. Wallace, B.C., Choe, D.K., Charniak, E.: Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference. pp. 1035–1044. Association for Computational Linguistics (ACL), Beijing, China (2015)
28. Yassine, A., Mohamed, L., Al Achhab, M.: Intelligent recommender system based on unsupervised machine learning and demographic attributes. Simulation Modelling Practice and Theory **107**, 102198 (2 2021). https://doi.org/10.1016/J.SIMPAT.2020.102198