Web Semantics: Science, Services and Agents on the World Wide Web 71 (2021) 100660

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

HeadlineStanceChecker: Exploiting summarization to detect headline disinformation

Robiert Sepúlveda-Torres *, Marta Vicente, Estela Saquete, Elena Lloret, Manuel Palomar

Department of Software and Computing Systems, University of Alicante, Apdo. de Correos 99 E-03080, Alicante, Spain

ARTICLE INFO

Article history:
Received 21 February 2021
Received in revised form 7 June 2021
Accepted 20 August 2021
Available online 27 September 2021

Keywords:
Natural Language Processing
Fake news
Misleading headlines
Stance detection
Applied computing
Document management and text processing
Semantic summarization

ABSTRACT

The headline of a news article is designed to succinctly summarize its content, providing the reader with a clear understanding of the news item. Unfortunately, in the post-truth era, headlines are more focused on attracting the reader's attention for ideological or commercial reasons, thus leading to mis- or disinformation through false or distorted headlines. One way of combating this, although a challenging task, is by determining the relation between the headline and the body text to establish the stance. Hence, to contribute to the detection of mis- and disinformation, this paper proposes an approach (*HeadlineStanceChecker*) that determines the stance of a headline with respect to the body text to which it is associated. The novelty rests on the use of a two-stage classification architecture that uses summarization techniques to shape the input for both classifiers instead of directly passing the full news body text, thereby reducing the amount of information to be processed while keeping important information. Specifically, summarization is done through Positional Language Models leveraging on semantic resources to identify salient information in the body text that is then compared to its corresponding headline. The results obtained show that our approach achieves 94.31% accuracy for the overall classification and the best FNC-1 relative score compared with the state of the art. It is especially remarkable that the system, which uses only the relevant information provided by the automatic summaries instead of the whole text, is able to classify the different stance categories with very competitive results, especially in the *discuss* stance between the headline and the news body text. It can be concluded that using automatic extractive summaries as input of our approach together with the two-stage architecture is an appropriate solution to the problem.

1. Introduction

Nowadays, disinformation and misinformation are two major problems that are increasing at great velocity [1] in pace with the exponential growth of information on the web and the need for robust verification methods. If handling this information overload is an arduous and complex task for both humans and machines, verifying its veracity has become a daunting yet unavoidable challenge. Both terms, misinformation and disinformation, allude to the inaccuracy and lack of veracity of certain information; however, while in the first case the delusion can be caused unintentionally, the latter actually seeks to deceive or misdirect deliberately [2]. In either case, they represent a type of phenomenon that, in the domain of digital news, can easily result in a massive confusion about the real facts, spreading on a viral scale. This is actually what the New York Times meant when they referred to a "Fake news" piece as a "made up story with the intention to deceive, often with monetary gain as a motive" [3].

The ideological and economic interests that potentially gain from this "information disorder" are the drivers of fake news. These interests aim to manipulate social opinion and reinforce preconceived opinions, thereby making people focus on thinking or acting in a specific way by, most of the time, appealing to their emotions rather than presenting the facts. This trend that has even prompted the advent and consolidation of a new term, "post-truth", which, according to the Cambridge Dictionary,[1] refers to "a situation in which people are more likely to accept an argument based on their emotions and beliefs, rather than one based on facts". For instance, this distorting phenomenon played an important role in *President Trump's election campaign 2016* [4] and *the Brexit referendum 2016* [5]. In the same way, business and commercial interests fabricate fake news to generate income through clickbait and misleading information. For instance, the National Report website, Disinformedia [6] or Victory

* Corresponding author.
E-mail addresses: rsepulveda@dlsi.ua.es (R. Sepúlveda-Torres), mvicente@dlsi.ua.es (M. Vicente), stela@dlsi.ua.es (E. Saquete), elloret@dlsi.ua.es (E. Lloret), mpalomar@dlsi.ua.es (M. Palomar).

1 https://dictionary.cambridge.org/.

https://doi.org/10.1016/j.websem.2021.100660

Lab [7] are examples of websites that produce and/or disseminate fake news.

Assessing the veracity of a news story is a complex task (either for expert journalists or for Artificial Intelligence). For this reason, the research community approaches the resolution of this task from several perspectives that imply different sub-tasks. In this manner, it is convenient to assess the veracity of a news story by splitting the task into simpler parts and dealing with them individually [8].

Following this approach, great attention and effort has been focused on the analysis and study of one of the most essential elements of a news item, its headline, in some cases focusing on the relationship between the body of the article and the headline, and in others considering the constitution of the headline itself. Headlines are fundamental parts of news stories, they summarize the article so that the reader clearly understands the content of the news story [9]. Nevertheless, the headline acts also as the prelude to the complete news story, and it should be written as an invitation for the reader to discover the full piece. A headline is therefore expected to be as effective as possible, without losing accuracy or becoming misleading [10].

In the scenario we have outlined, where the information stream is permanently growing and filtering content can be overwhelming, the role of headlines is crucial. On the one hand, an appropriate headline can help us to identify the content of most interest to us, but on the other hand, and due to this data deluge, it can be tempting to read only the headlines and share the news feed without having read the entire story. Consequently, stories can often go viral because of an attractive headline despite the lack of true information in the body text. This phenomena manipulates public opinion and affects the credibility of social media [11,12]. In particular, the research conducted in [11] found that 59% of the URLs mentioned on Twitter were not clicked at all. This suggests that people are more willing to share an article than access and read it, so they directly read and share the headline (and link), without making the effort to go deeper and check it. Considering this, the headline of a news article should faithfully summarize the body text, without including deception or misinformation, in order to maintain accuracy and veracity of the entire article.

Unfortunately, in practice, headlines in digital media tend to be more focused on attracting the reader's attention (with little regard for accuracy) thus leading to mis- or disinformation through erroneous/false facts or headline/body dissonance [13]. In this context, headlines can be classified into two classes [14]:

- **Clickbait headlines**: Clickbait refers to content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page with the purpose of monetizing the "views" through advertising revenue (the more clicks, the more money earned). This type of headline is often ambiguous and exhibits a particular writing style to directly exploit human curiosity, for instance by using exclamatory or interrogative headlines that urge audiences to click on the link to discover the missing information [14]. Typically, clickbait headlines are spread on social media in the form of short teaser messages that may read like the following cited examples:
  - *"Man tries to hug a wild lion, You won't believe what happens next!"*[2]
  - *"The first lady of swearing! How a ten-year-old Michelle Obama lost out on a 'best camper' award because she wouldn't stop cursing"*.[3]

  Existing methods for automatically detecting clickbait headlines usually treat the task as a classification problem (clickbait/non-clickbait), and exclusively focus on the headline (its writing style or structure) rather than considering the content of the news itself [13,15].
- **Misleading headlines**: Headlines thus classified significantly misrepresent the findings reported in the news article [16], by exaggerating or distorting the facts described in the news article. The reader can only discover the inconsistencies after reading the news body text [14]. Although in the literature these headlines are sometimes referred to as *incongruent headlines*, in this work we will refer to them as *misleading headlines* since the term represents a more comprehensive concept.

  Some important nuances that are part of the news body text are missing in the headline, causing the reader to come to the wrong conclusion. In contrast to clickbait headlines, the language used does not necessarily incite the reader to click on it, but it is designed to trigger emotion or excitement [16].

  Examples of misleading headlines are shown below (also reported in [13,16], respectively):
  - *"Ebola in the air? A nightmare that could happen"*.[4]
  - *"Air pollution now leading cause of lung cancer"*.[5]

  In order to automatically detect misleading headlines, the news body text must be analyzed to extract the evidence from which the headline has been derived, thereby detecting the headline/body text discrepancy in the absence of such evidence. The task of identifying the relation between a headline and the news article it refers to has been addressed in recent research (see Section 2) as a stance detection problem. This type of approach involves estimating the relative perspective, namely the stance, of one piece of text, such as a claim or a news article, towards another, for example, a topic, a statement or a headline [17].

In the context of headline/body text dissonance, the main objective of this research is to propose an approach that relies on semantics and deep learning techniques to automatically determine the stance of the headline with respect to its body text. By this means, the problem of misleading headlines can be addressed. The approach is hereafter referred to as *HeadlineStanceChecker*. Given a news headline and its corresponding body text, our proposal assigns the headline one of these four classes (*unrelated*, *agree*, *disagree* or *discuss*), indicating the headline stance, and validating and checking whether the headline is faithfully reflecting the information provided in the news article.

The most interesting aspect of solving misleading headline detection as a stance detection task is that it is not only focused on determining whether or not a headline is consistent with its body text, but it is also a fine-grained classification that determines the type of dissonance involved.

We explore the treatment of this task as a two-stage neural classification problem in which only the essential information of the news item is processed, rather than the whole news item. We therefore use the summaries because besides containing the key information of the news story, we hypothesize that the abridged version will not only increase task efficiency but also that of the neural models. Neural models can have a negative impact on efficiency when processing long texts, so previous studies either used the first sentence of the text [18] or a specific fragment [19] to combat this problem. Therefore, the use of text summarization,

[2] https://bit.ly/2FEddK2 (accessed online 15 February, 2021).

[3] https://www.dailymail.co.uk/news/article-3004975 (accessed online 15 February, 2021).

[4] https://cnn.it/2NeuNZj (accessed online 15 February, 2021).

[5] https://bit.ly/2Tajoxx (accessed online 15 February, 2021).

which, to the best of our knowledge, has not been previously exploited for stance detection, could be beneficial in this context.

To summarize, the main novelties of *HeadlineStanceChecker* are twofold:

- the adoption of a divide-and-conquer strategy by proposing a two-stage neural classifier for performing the headline stance task; and
- the use of summarization techniques based on Positional Language Models (PLM). These models leverage semantic knowledge to detect the evidences and essential information within the news article so as to generate automatic summaries that will be used as substitutes of the full body text for the whole classification process. We expect this approach to be more efficient in dealing with the headline stance classification task.

The paper is structured as follows: Section 2 presents the related work regarding misleading headlines, as well as a brief review of the state of the art in text summarization; Section 3 presents our proposed architecture for *HeadlineStanceChecker* (explaining each of the stages in detail); Section 4 describes the experiments carried out and the evaluation environment; Section 5 reports and discusses the results of the proposed approach (comparing them to other competitive systems); and , Section 6 presents conclusions and outlines the main direction for future work.

2. Related work

HeadlineStanceChecker has been conceived as an automatic method to classify a news story in terms of the relation between its body text and its headline. The main motivation of developing such approach is to provide a tool that helps both professionals and readers to identify misleading or fraudulent media and information, thus preventing harmful consequences.

Fake news research has opened up an immense field of work that encompasses multiple areas and approaches. Both linguistic and non-linguistic aspects are being studied, so that elements as diverse as image verification, analysis of reputation and authorship, or the network dissemination patterns of misleading stories fall within its field of interest. For brevity, we focused on the research directly related to our proposal, but comprehensive studies can be found in [8,20,21].

Therefore, in this section, first we present an overview of recent work done in Stance Detection and, next, an in-depth review of the existing detection strategies for misleading headlines is conducted. Finally, given that one of the novelties of the paper is using summarization techniques leveraging essential information to characterize headlines, a brief review is presented of the state of the art in text summarization.

- **Stance Detection Overview**
  From an overall perspective, stance detection can be defined as the task of identifying the perspective of an author or text against a given target in the form of one topic, claim, headline or even a personality [22,23]. Hence, there exists a tuple of elements (the text on the one side, the target on the other side) and a classification process shaped to determine how the former stands towards the second: does the text support the topic? does it disagrees with the claim? The names of the classes (e.g. *support*, *against*, *for* or *neutral*) depend on the precise problem. The task, which concerns a diverse range of domains, is studied in such varied areas as political debates [24,25], student essays [26], online forum debates [27] or even internal company discussions [28,29].
  A great deal of work in opinion mining has been devoted to detect the stance of tweets or other types of short texts as rumors [30] or microblogging statements. Examples of targets posed in the available datasets could be "Hillary Clinton" for personality, "Atheism" as a particular topic or the claim "E-cigarettes are safer than normal cigarettes". Shared tasks offering such datasets and fostering the research on the matter have arisen in different languages. SemEval-2016 posed the sub-task for detecting stance in tweets [31], providing around 5 thousand tweets in English covering five commonly known topics. The task has inspired numerous approaches that develop either traditional proposals (e.g. K nearest neighbor [32], Support Vector Machines [33] or latent features provided by methods such as Latent Dirichlet Allocation [34]); or those inspired by neural network frameworks, by using, for example, bidirectional conditional encoding [35], bidirectional Long Short-Term Memory neural networks [36] or Attention based Convolutional Neural Networks [37]. Besides, there are available public datasets that support the development of new interesting work, such as the *Multi Perspective Consumer Health Query dataset* [38] dedicated to detecting the stance of sentences collected from quality articles towards five different claims (e.g., "Sun exposure causes skin cancer"). In [23], an in-depth study on different approaches to the two tasks mentioned above can be found. Regarding languages other than English, the necessity for well-annotated data led to the proliferation of both annotation efforts and shared tasks aimed to advance research, such as *StanceCat*, presented at IberEval 2017 as a stance detection task for tweets in Spanish and Catalan [39], a proposal and a dataset of short messages in Russian internet forums [40], or even projects combining a larger number of languages (French, Italian, Spanish, English) [41,42].
  In contrast to such approaches, research on stance detection based on longer documents, as in the current scenario, faces different challenges. Dealing with discourse, as a coherent and cohesive set of sentences, adds a certain complexity not present when processing shorter utterances. Within the discourse, an argument may develop in such a way that some sentences may show support for the claim, while others may seem to deny it, and only by considering the document as a whole can the stance be effectively identified. It is in this context that *HeadlineStanceChecker* has been developed, and next, we introduce the related work concerning the specific task.
- **Misleading headlines**
  The task of detecting misleading headlines for the present research involves classifying the stance of the article body with respect to the claim made in the headline into one of the following four classes: (a) *agrees* (agreement between body text and headline); (b) *disagrees* (disagreement between body text and headline); (c) *discusses* (same topic discussed in body text and headline, but no position taken); and, (d) *unrelated* (different topic discussed in body text and headline).
  This task (headline stance detection) quickly emerged in the context of fake news analysis, triggered by a demand for new technologies to prevent and combat the phenomenon, together with an increase in the availability of annotated corpora [8]. In this context, research challenges and competitions were proposed. The most recent and important ones are next reviewed in detail.
  The *Fake News Challenge*[6] (FNC-1) [43] was created using Emergent dataset [17] as a starting point (this dataset has been extracted from the Emergent Project [44], a rumor

[6] http://www.fakenewschallenge.org/ (accessed online 15 February, 2021).

debunking project). FNC-1 aims to compile a gold standard to explore Artificial Intelligence technologies, especially ML and Natural Language Processing (NLP), applied to detection of fake news. To carry out this macro-challenge, the organizers decided to start with stance detection. In this case, the FNC-1 dataset was released, with around 75,000 instances that were classified as follows: *agree*, *disagree*, *discuss* and *unrelated*.

For example, given the headline *"Robert Plant Ripped up $800M Led Zeppelin Reunion Contract"*, the following fragments[7] would illustrate the different classes mentioned, according to the gold-standard annotations in the FNC-1 dataset:

- ***Agrees***: The body text agrees with the headline. Example evidence:*"[...] Led Zeppelin's Robert Plant turned down 500 MILLION pounds to reform supergroup"*.
- ***Disagrees***: The body text disagrees with the headline. Example evidence:*"[...] No, Robert Plant did not rip up a $800 million deal to get Led Zeppelin back together"*.
- ***Discusses***: The body text discusses the same topic as the headline, but does not take a position. Example evidence: *"[...] Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal"*.
- ***Unrelated***: The body text is not related with the headline. Example evidence: *"[...] Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today"*.

The FNC-1 competition received a total of 200 submissions achieving relative scores[8] of around 82% in the best ranked submissions. The organization proposed a simple baseline using hand-coded features and a gradient boosting classifier, available at Github.[9] The three best systems in this competition were Talos [45], Athene system [46] and UCLMR [47] in this order. Talos [45] applied a one-dimensional convolution neural networks (CNN) on the headline and body text, represented at the word level using Google News pretrained vectors. The output of this CNN is then sent to a multi-layer perceptron (MLP) with 4-class output: *agree*, *disagree*, *discuss*, and *unrelated*, and trained end-to-end. Using this combination CNN-MLP, the system outperformed all the submissions and achieved the first position in the FNC-1 challenge.

Recently, other works used the FNC-1 for their experiments and the performance obtained in the competition improved. For instance, [48] addressed the problem proposing a hierarchical representation of the classes, which combines agree, disagree and discuss in a new related class. A two-layer neural network is learning from this hierarchical representation of classes and a weighted accuracy of 88.15% is obtained with their proposal. Furthermore, [49] constructed a stance detection model by performing transfer learning on a RoBERTa deep bidirectional transformer language model by taking advantage of bidirectional cross-attention between claim-article pairs via pair encoding with self-attention. They reported a weighted accuracy of 90.01%.

Outside the FNC-1 Challenge and dataset, there is other research that also addresses the stance detection tasks, determining the relation of a news headline with its body text. Some authors extracted key quotes [50] or claims [51] to facilitate the detection. There is also work related to argument mining analysis, in which the headline represents an argument that is not supported by claims in the text. Moreover, in addition to using argument mining for solving stance detection, this problem could benefit from other tasks which detect semantic relations within the text, such as contradiction [52], contrast [53] and entailment [54].

- **Text Summarization**
  Previous research in Text Summarization has been shown to have a positive impact on society since the use of summaries has been beneficial in different areas, such as education (where summaries are used to support reading comprehension tasks [55–58]) business, by producing, for instance, an automatic summary of event logs to help analysts [59], or health, regardless of whether the summaries were created manually [60,61], or automatically [62]. This is partly due to the capability of summarization methods to identify the most relevant information of a document, and condense it into a new text, thereby helping to reduce time and resources when it comes to manage large amounts of data. These methods have proven to be effective when integrated as an intermediate component of more complex systems.
  The journalism field, and specifically the news domain, has been one of the most representative areas in which summarization has traditionally focused from the outset, partly thanks to the development of appropriate corpora (e.g. DUC, Gigaword, CNN/DailyMail) [63], and the wide range of techniques and approaches to help digest this type of information [64–67]. Besides the various summarization types that have been developed for this domain (single-document, multi-document, extractive, abstractive, generic, topic-oriented, etc.), there is a significant amount of research on the task of headline generation using summarization techniques [68–70], and more recently using Deep Learning [71–73]. However, none of them have exploited either the headline or the summarization techniques as an intermediate stage to further extract the semantic relationship between the headline and the news body text, and detect possible incongruities to fight against the fake news problem.

Although summarization has been used for fake news detection [74,75], as well as in the context of online discussions and social media to detect whether the author of a comment is in favor of or against a given target (e.g. entity or topic) [76,77], to the best of our knowledge, summarization has not been directly applied to the stance detection problem of misleading headlines, as proposed in this study.

Summarization was mentioned as a potential effective methodology for dealing with the problem of incongruent headlines in [78], but from a different perspective, which involved using headline generation to create a new headline that could be then compared to the existing headline by measuring the distance between them. More recently, an updated comprehensive survey concerning the stance detection task [79] shows that there is a lack of research where summarization is applied to this task, although a new type of summarization, called stance summarization, is outlined. However, stance summarization involves the generation of a new type of summary which includes a stance, but it is not comparable to the approach presented in this paper as the summaries are not incorporated into the stance detection process.

In another survey, conducted by [80], the authors compile the available information regarding existing research addressing this problem, and only the work of [81] summarized the news body into a single sentence to be compared to the given claim and determine its overall veracity, an approach which aligns

7 Examples extracted from Fake News Challenge website fakenewschallenge.org.

8 Measure score used in the Fake News Challenge competition.

9 https://github.com/FakeNewsChallenge/fnc-1-baseline (accessed online 15 February, 2021).

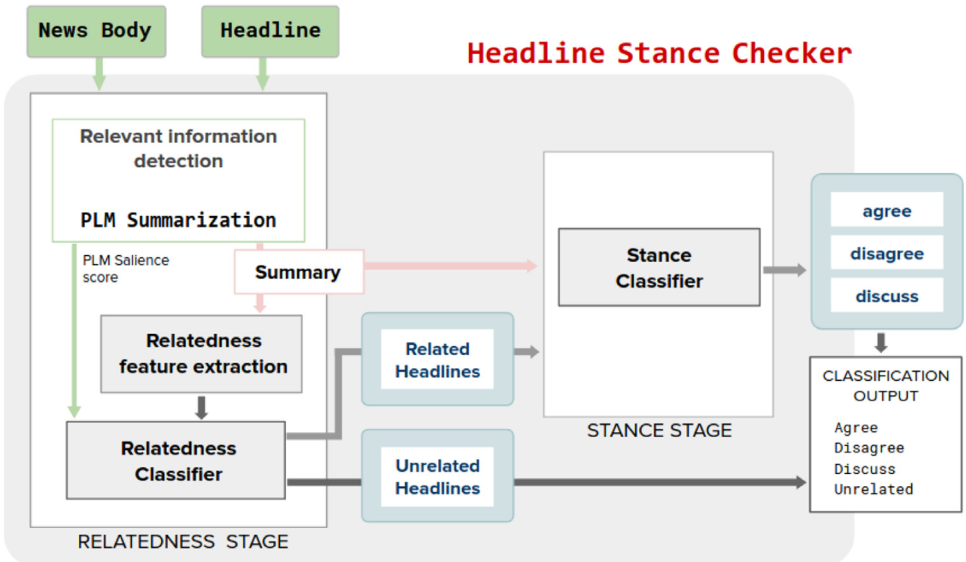

Fig. 1. *HeadlineStanceChecker* architecture.

with that suggested in [78] as aforementioned. By contrast, our research goes beyond summarizing the whole document into just one sentence, and provides a summary that could be acted as a substitute of the whole body text.

The *HeadlineStanceChecker* proposal is based on the fact that semantic information and discourse structure are captured through PLMs which, in turn, are exploited as a summarization technique. PLMs allow key spots and relevant information to be located in the news body text, and they are then used to create a summary of the news. By this means, the news article is reduced to its essential information, which is then compared to its headline. Our proposed model to detect misleading headlines, by relying on their stance towards the article's content, directly uses this summary of the news instead of the whole news body text, enabling a more accurate comparison to its headline.

3. HeadlineStanceChecker architecture

The HeadlineStanceChecker approach involves two-stages (see Fig. 1), thus addressing the task as a two-level classification problem. The first level corresponds to a *Relatedness Stage*, while the second corresponds to a *Stance Stage*. An additional novelty is the use of summaries generated in the first stage for the whole process instead of the full body text (i.e., the *Relatedness Stage*).

In this manner, given the inputs, namely the candidate headline and the news article body text, a summary of the news body will be created in the *Relatedness Stage* to later determine the headline's stance regarding the news article as either *related* or *unrelated*. Afterwards, in the *Stance Stage*, the examples classified as *related* in the previous stage, are further classified into three possible values: *agree*, *disagree*, or *discuss*.

A more detailed description of both stages and the different modules involved in performing the stance classification is provided here-under.

3.1. Relatedness stage

The *Relatedness Stage* will determine whether the headline is classified as *related* or *unrelated* with respect to the body text of the news article. The inputs of this stage are both the text body and the headline, resulting in a binary classification. The outputs of this stage are:

- The *headlines* classified as *related* or *unrelated*.
- The *summary* of the news content, obtained in a relevant information detection module.

To produce the above outputs, three modules are proposed: (i) relevant information detection; (ii) relatedness feature extraction; and, (iii) relatedness classification.

3.1.1. Relevant information detection module

The relevant information detection module aims to create a summary revealing the important information of the input news article in relation to its headline.

The task of summarization has generally been carried out from a statistical perspective that only considers the elements of the text with no regard to their structure (or in those cases where the structure is taken into account, it is already known beforehand, such as in the case of scientific articles). Conversely, PLMs represent a type of statistical language model that allows information to be considered by taking into account both the relevant elements of the text and also their location in the document. They define a dynamic method for detecting key aspects of the text independently of the domain and textual genre to which it is related. Besides, PLMs have proved valuable in other areas such as information retrieval [82] and language generation [83].

From the semantic perspective of the text, its essence can be more effectively captured and synthesized by considering the document not as a mere sequence of sentences, but as a coherent and cohesive source of meaning, traversed by semantically related entities and actions. Considering this, we chose PLMs as the cornerstone of our module given that they can be configured to include the identification of named entities within the story, together with the representation of the words as synsets (sets of synonyms accounted under an identifier), allowing a further abstractive step on the basis of Wordnet [84], a hierarchical database of semantic relations. Consequently, PLMs help to incorporate both the semantics derived from the relevant lexical units together with the meaning derived from the text as coherent discourse. Previous studies demonstrated that PLMs were suitable for summarization tasks [85] and, moreover, a preliminary research was conducted analyzing and comparing different summarization methods for the stance detection task (including extractive, abstractive and hybrid ones) also showing that PLM-based summarization yielded the most stable results [86].

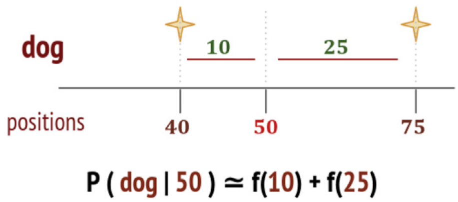

Fig. 2. Example of the type of computation performed to obtain the value of the PLM for the position 50 regarding the element *dog* of the vocabulary.

PLM essentials. Fundamentally, the PLMs state that for every position i within a document D it is possible to calculate a score for each element w that belongs to the document's vocabulary. The decision as to the kind of elements that compose the vocabulary is made when designing the module. The calculated score displays the relevance of each element w in every precise position i, based on its distance to other occurrences of the same element throughout the document. The score is higher when the neighbor element is closer within a scope to compute the value that goes beyond the sentence limits, taking into account the whole document. In order to express the distance to the occurrence of the entity in the neighborhood, a propagation function $f(i,j)$ is applied.

Eq. (1) defines how the score for word w in position i is computed:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \quad (1)$$

where $|D|$ refers to the length of the document, V is the vocabulary, $c(w,j)$ indicates the presence of element w in the position j, and $f(i,j)$ is the propagation function that rates the distance between i and j. In this case, and taking as bases previous work on the matter [83], a Gaussian kernel is adopted as the propagation function.

Fig. 2 illustrates the idea behind the PLM reasoning.

PLM for summarization. The manner in which PLMs are employed to perform the summarization task comprises three stages. First, we need to conduct the definition of the vocabulary as a parameter for the PLM module. In our current configuration, the vocabulary is composed of the synsets corresponding to nouns, verbs and adjectives, together with the named entities that appear along the text. In order to get this semantic information, we use Freeling [87], an open source tool that allows linguistic analysis with different levels of granularity.

From this stage, a representation of the text that involves both the vocabulary and the positions of its elements is obtained.

Second, we create a seed, i.e., a set of words that can be significant for the text and will help the system to discard irrelevant parts of the discourse. The given headline is taken as seed in our configuration. It needs to be analyzed with the same tools as the source text (Freeling). As a result, a second vocabulary is then built from it.

Finally, the processing of the PLM against the seed allows us to compute scores for the text elements that are now conditioned by the information in the headline. At this stage, we have already calculated a collection of values associated with every relevant element using the PLMs for each position of the text. The aggregation of the different values related to each of those positions results in a vector with same length of the document, the Score Counter (SC), so that the score held in the index i will express the value for the position i in the text. Those positions in the text that show local maximums in the SC are retrieved as the most relevant points of the document. The sentences to which these positions belong are then selected as candidates for the summary. Since a value has been calculated for each position in the sentence, we can obtain a score for the sentence itself:

$$S_{score} = \sum_{i \in S} SC[i] \quad (2)$$

with S representing the sentence, and i indicating the positions within the document for that sentence.

These values also allow us to select from the candidates the sentences that will constitute the news extractive summary. Moreover, the computed values are necessary to define a new feature, named **PLM Salience Score**, which will be used for the relatedness classifier in the next step. Its value derives from the aggregation of each score S_t associated with each sentence t included in the summary, following Eq. (3). Let S^* represent the set of the sentences belonging to the summary, the *PLM Salience Score* for a document D would be calculated as:

$$PLM\ Salience\ Score_D = \sum_{t \in S^*} S_t \quad (3)$$

3.1.2. Relatedness feature extraction

Besides the relevant information (i.e., the summary) and the PLM Salience Score obtained in the previous module, two similarity features are used as input to the relatedness classifier applied next. To obtain the features, the headline and the generated summary are used. They are described next:

- **Cosine similarity**: The cosine similarity between headline and summary of body text is computed. This feature is used to measure how similar the headline and summary are, irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space [88]. The cosine similarity is advantageous because even if the two similar documents are far apart by Euclidean distance (due to the size of the document), chances are that they may still be oriented closer together. The smaller the angle, the higher the cosine similarity [89]. Although this metric is relatively basic [90], it usually brings significant improvements to retrieval models [88]. The cosine similarity measure between two vectors X and Y is obtained following Eq. (4) [91]:

$$Cosine\,similarity\,(|X, Y|) = \frac{x \cdot y}{\|x\|\ \|y\|} \quad (4)$$

where $x \cdot y = \sum_{i=1}^{n} x_i y_i$ and $\|x\| = \sqrt{x \cdot x}$

For the calculation of cosine similarity, the text pairs are converted into Term Frequency–Inverse Document Frequency (TF–IDF) vectors, using the tools provided by scikit-learn [92].

- **Overlap coefficient**: This feature is defined as the intersection between two sets A and B. In the current scenario, these sets contain the ngrams belonging either to the headline or the summary [93]. The overlap coefficient is given by Eq. (5) [94]:

$$Overlap\ coefficient\,(A, B) = \frac{A \cap B}{\min(|A|, |B|)} \quad (5)$$

If set A is a subset of B or the converse, then the overlap coefficient is equal to 1 else overlap coefficient should be between 0 to 1 [95].

3.1.3. Relatedness classification

The relatedness classification module exploits the PLM Salience Score, the relatedness features previously processed, as well as the summary to finally classify the headlines as *related* or *unrelated*. The proposed architecture is flexible to choose any model that allows classifiers to be improved.

In this case, the design of the relatedness classification module is based on fine-tuning the RoBERTa (Robustly optimized BERT approach) pre-trained model [96], applying a classifier to its output afterwards.

First, the headline and the summary are concatenated and processed with the RoBERTa model. The resulting vector is consecutively multiplied by the three features (PLM Salience Score, Cosine similarity, Overlap coefficient) to finally carry out the classification using a Softmax activation function in the output layer.

Specifically, we have chosen RoBERTa Large model (24 layer and 1024 hidden units) since it achieves state-of-the-art results in General Language Understanding Evaluation (GLUE) [97], Reading Comprehension Dataset From Examinations (RACE) [98] and Stanford Question Answering Dataset (SQuAD) benchmark. Similar to [49,96,99], in this work we fine-tune RoBERTa to efficiently address a task that involves comparing sentences. RoBERTa optimizes Bidirectional Encoder Representations from Transformers (BERT) [100] by adding several modifications but without altering the original architecture, an approach that improves the results with respect to BERT in the main NLP tasks [96]. Some of those modifications involve: eliminating the prediction of the next sentence; performing the training on a greater volume of data; enlarging the batch size; and, lengthening the input sequence.

To create the classifier, the *Simple Transformers library*[10] was used, which creates a wrapper around *HuggingFace's Transformers library* for using Transformer models [101]. *Simple Transformers* is an NLP library that allows the modification of hyperparameters so as to train, evaluate, and make predictions using the best state-of-the-art models.

In our model, the hyperparameter values are: maximum sequence length of 512; batch size of 4; training rate of 1e-5; and, training performed for 3 epochs. These values were established after successive evaluations, following previous experiments on this model [49,96,99].

3.2. Stance Stage

Given the related headlines obtained through the first stage on the proposed architecture, the main goal of this stage is to determine their type considering the remaining stances: *agree*, *disagree* or *discuss*. Therefore, the claim made in the headline can be finally classified into one of three classes left.

The inputs of this stage are:

- The *headlines* classified as *related*.
- The *summary* of the news content.

The output of this stage then is the final classification of the related headlines, where each of them is assigned one of the following possible stance values: *agree*, *disagree* or *discuss*. These classified headlines together with the *unrelated* headlines determined before, will comprise the final output for the whole *HeadlineStanceChecker* approach.

[10] https://simpletransformers.ai/ (accessed online 15 February, 2021).

Table 1
Description FNC-1 dataset considering number of documents.

	Documents	Headlines	Instances
Train set	1,683	1,683	49,972
Test set	904	904	25,413
Complete dataset	2,587	2,587	75,385

3.2.1. Stance classification

As in the *Relatedness Stage* (Section 3.1.2), the extractive summary generated in Section 3.1.1 is also used here.

Similar to the Relatedness classification module, this stage has been built using RoBERTa as the selected model capable of improving the classification. In this case, no additional features are considered, only two dense layers are included to reduce dimensions and, finally, the Softmax classification layer. The hyperparameters of the model used in this classifier are the same as those of the Relatedness classification, except for the classification output which in this case is of three classes: *agree*, *disagree*, *discuss*.

4. Experiments and evaluation environment

The proposed approach was applied and evaluated in the context of the Fake News Challenge FNC-1 whose goal was to determine a headline's stance by classifying it in relation to its body text into 4 classes: *unrelated*, *agree*, *disagree*, and *discuss*. In this section, we first describe the corpus provided in this challenge. Second, we explain the experiments performed with different configurations of our system. Finally, the evaluation metrics used are outlined. The results obtained will be presented, discussed and compared to the other participating systems in the challenge in subsequent sections.

4.1. Fake News Challenge Dataset

The experimentation is conducted over the FNC-1 dataset whose instances are labeled as *agree*, *disagree*, *discuss* and *unrelated*.

The dataset was split into a training set (66.3%) and a testing set (33.7%), where neither the headlines nor the body text overlapped. The distribution of documents (bodies and headlines) is presented in Table 1.

As the distribution of the classes indicates in Table 2, there is a significant imbalance for both the training and testing sets where the instances of the *unrelated* class alone (over 70%) are greater than the sum of the remaining classes. At the other extreme, the *disagree* class is remarkably lower compared to the others.

4.2. Experiments

To measure our system's performance, a set of experiments was conducted as follows, the results of which will be shown and discussed in Section 5. Our experiments can be replicated at Github[11]:

- *Relatedness Stage Validation*: The aim of this experiment is to assess the performance of this classification stage, where *related* or *unrelated* headlines are initially identified. First, we analyze and compare the performance of the classifier when either summaries or the full body is employed. Second, we conduct an ablation study to verify whether the relatedness features used for the classifier make a positive contribution.

[11] https://github.com/rsepulveda911112/HeadlineStanceChecker (accessed online 10 september, 2021).

Table 2
Distribution of FNC-1 dataset stances.

	Agree	Disagree	Discuss	Unrelated
Train set	3,678 (7.36%)	840 (1.68%)	8,909 (17.82%)	36,545 (73.13%)
Test set	1,903 (7.48%)	697 (2.74%)	4,464 (17.56%)	18,349 (72.20%)
Complete dataset	5,581 (7.4%)	1,537 (2.03%)	13,373 (17.73%)	54,894 (72.81%)

Table 3
Relatedness classification results: class-wise F_1 Score and F_1m using automatic summaries vs. full news text.

System	F_1 Score		F_1m
	Related	Unrelated	
Relatedness Stage FNC-1-Summary	**98.38**	**99.40**	**98.89**
Relatedness Stage FNC-1-Body-text	98.36	99.37	98.86

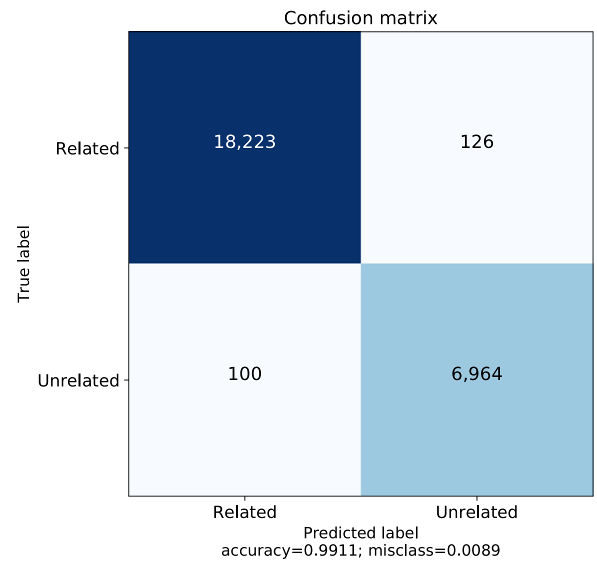

Fig. 3. Confusion matrix resulting from the *Relatedness Stage FNC-1-Summary*.

- *Stance Stage Validation*: The goal of this experiment is to determine how accurate the *Stance Stage* is when the errors produced by the *Relatedness Stage* are avoided, thereby using an ideal input for this stage, i.e., the gold-standard headlines annotated as *related* in the FNC-1 corpus. By this means, we can measure the effectiveness of this stage in isolation. Furthermore, to validate the extent to which our proposed stance detection model can be generalized, we apply it to a different headline stance detection dataset, i.e. Emergent dataset [17].
- *HeadlineStanceChecker Validation*: In this last experiment, the entire system (integrating the Relatedness and Stance classifiers as a two-step classifier and using summaries as input for the whole process instead of the full text) is tested. Its performance is then compared to other configurations of the model as well as to competitive state-of-the-art systems.

In addition, we also investigate the system performance considering two different inputs: summaries and full body. This experiment and its further analysis is detailed in Section 5.4.

4.3. Evaluation metrics

Originally, the organizers of the FNC-1 challenge proposed the *Relative Score* metric, which assigned a higher weight to examples correctly classified, as long as they belonged to a different class from the *unrelated* one. The rationale behind this metric was to address the highly imbalanced distribution of the classes caused by the over-representation of the *unrelated*.

However, as pointed out in [102], the inner imbalance among the three *related* classes (*agree*, *disagree*, and *discuss*) was not addressed. Therefore, following [102], this study incorporates, in addition to the FNC-1 relative score, both a measure of F_1 class-wise and a macro-averaged F_1 (F_1m) as the mean of those per-class F scores so as to address the imbalance among the less represented classes. The advantage of this measure rests in it not being affected by the size of the majority class. Additionally, average accuracy is also obtained.

5. Results and discussion

This section presents the results obtained in each of the experiments described in Section 4.2. The values are expressed in percentage mode.

5.1. Relatedness Stage Validation

Our first experiment was designed to evaluate the first module as an isolated element of the system, acting as a binary classifier. In this case, we were not evaluating the system to detect *agree*, *disagree* or *discuss* examples, but to perform *related* versus *unrelated* classification. We carried out an analysis of the classification results and also an ablation study that considered the following involved features: cosine similarity; PLM Salience Score; and, overlap coefficient.

The performance of the relatedness classifier was first validated by analyzing whether the use of summaries had a positive impact on the output compared to using the whole document. The results are shown in Table 3. Both approaches used the three features previously described in Sections 3.1.1 and 3.1.2.

Relatedness Stage FNC-1-Summary refers to an experiment that uses summaries both to calculate features and to enter the classification model, whereas the *Relatedness Stage FNC-1-Body-text* approach uses the body text instead of the summary as input to the relatedness stage to classify the headline. The results validate the use of summaries as a useful approach to the stance detection problem as even if some information is excluded, the findings indicate a slight improvement when using the summarized text.

The approach that uses summaries throughout the process is able to improve the related class, which is the minority class. Figs. 3 and 4 show each confusion matrix of the two approaches with minimal variation in the classification, thus showing that the use of summaries does not harm the results of this classifier.

These results show that, by using the PLMs to condense the relevant information from a piece of news, the resulting summaries offer an attractive substitute for the full news text, enabling a reduction of the computational load for the classifiers, which increases when dealing with longer texts.

Furthermore, to evaluate the influence of the added features in the relatedness stage, an ablation study of the features extracted from the summary has been conducted. Each feature (Cosine similarity, PLM Salience Score and Overlap coefficient) has been removed and an experiment has been designed that will return the results of the classification without the incidence of the removed feature. To the extent that the classification result

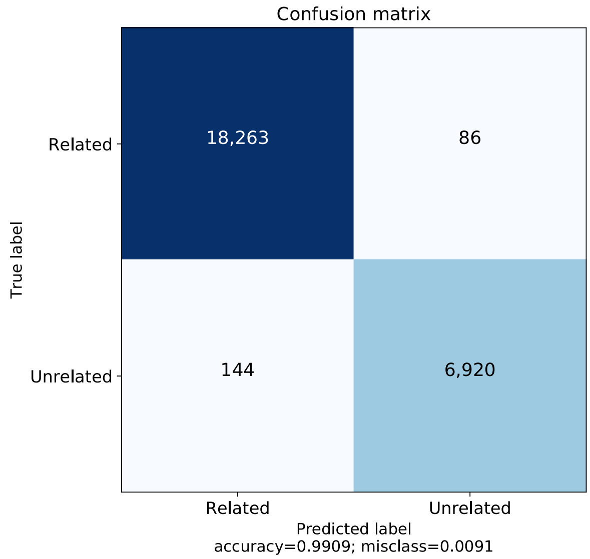

Fig. 4. Confusion matrix resulting from the *Relatedness Stage FNC-1-body*.

Table 4
Ablation study results for the features used in the Relatedness Stage. To facilitate reading and comparison, we have also included the non-ablation results.

Removed feature	F_1 Score		F_1m
	Related	Unrelated	
Cosine similarity	98.24	99.32	98.78
PLM Salience Score	98.00	99.23	98.61
Overlap coefficient	98.10	99.27	98.68
Non-ablated results	98.38	99.40	98.89

is worse, this would imply that the eliminated feature has a great influence on improving the classification results. The most influential feature for the classification was observed to be the PLM Salience Score as the experiment that does not use the PLM score obtains the worst results, followed by the one that does not use overlap coefficient and, finally, by the one that uses cosine similarity. Table 4 shows the ablation study.

5.2. Stance Stage Validation

This experiment was designed to determine the validity of the *Stance Stage*. This task can be tackled as a double question, since two fundamental issues arise: (i) the validity of the *Stance Stage* as a general proposal; and (ii) the effectiveness of the *Stance Stage* performance within an ideal case.

As for the first issue, this experiment aims to demonstrate that the approach is not an ad-hoc solution but a general one. For this purpose, the *Stance Stage* was applied to a different stance dataset called Emergent dataset[12] [17]. For this dataset, each example results from a combination of one article and its headline, and one claim. There are three different options for describing how a claim can be related to a piece of news. Specifically, each example was manually labeled by a journalist with one of the following tags: *for*, if the article states that the claim is true; *against*, if it states that the claim is false and *observing*, when the claim is reported in the article, but without assessment of its veracity. The dataset is composed of 2,595 examples, derived from the combination of 2,571 news, 2,536 headlines and 300 claims (see Table 5 for further details of the dataset).

To replicate our experimental environment with this dataset, the equivalence between labels in both datasets regarding their meaning is *for* $\simeq$ *agree, against* $\simeq$ *disagree and observing* $\simeq$ *discuss*.

Therefore, to validate the generalization of the approach, Table 6 includes the following performance results:

- *Emergent Upper Bound*: This experiment is performed as an upper bound by using a human-written headline created by a journalist, and considering it as a perfect summary that comprises the main information of the news body text. Nevertheless, this upper bound is only applicable to the Emergent dataset since in the case of FNC-1 no journalist-written headline is provided, as occurs in the case of the Emergent dataset.
- *Stance Stage using Emergent Dataset*: Our model is trained with the Emergent dataset and the Stance Stage is applied to it.
- *Stance Stage tested with Emergent, but trained with FNC-1*: This performance uses the Emergent dataset to test the Stance Stage but with the model trained on the FNC-1 so as to demonstrate the extent to which our proposal can be generalized.

The second aspect that needs to be addressed relates to the appropriateness of this second stage and its performance by isolating this stage from the rest of the system. The strategy here is focused on avoiding the errors inherited from the previous stage. To achieve this, only the examples tagged as *related* from the FNC-1 Gold-Standard are used and evaluated. The results of this performance correspond to *Stance Stage FNC-1* row in Table 6.

The analysis of the results obtained in this stage regarding the comparison of the performance using Emergent dataset are very promising considering that this model is using automatic summaries. The results are very close to the upper bound obtained by using human-made summaries. Analyzing per class, using the Emergent dataset for a training and testing task, the disagree class is even better classified by using automatic summaries. Additionally, when the approach is trained on the FNC-1 dataset and the test is carried out on the Emergent dataset, the discuss class surpasses the upper bound.

Regarding the performance of the Stance Stage in isolation, i.e., without considering the *Relatedness Stage*, the results present a slightly better performance than the whole approach with an increase of 2 percentage points (see Table 7). This was to be expected since errors derived from the *Relatedness Stage* are avoided. To conclude, these figures demonstrate that the approach, apart from potentially being a general solution, also demonstrates that using summarization of the body text as input is useful for the stance detection task, since the performance is very close to the upper bound proposed at Emergent.

5.3. HeadlineStanceChecker validation

The results of the *HeadlineStanceChecker* are shown in Table 7. This approach integrates the Relatedness and Stance classifiers and only uses automatic summaries for these two classifiers (but for the Relatedness classifier, the external features are included). This table contains the performance for the class-wise F_1, macro-average F_1m, accuracy (Acc.) and the relative score (Rel. Score). Moreover, it also provides the results obtained by competitive state-of-the-art systems together with additional configurations that were also tested.

The 3 first rows are the top-3 best systems that participated in the FNC-1 challenge. The results for each of the evaluation metrics

[12] https://github.com/willferreira/mscproject/ (accessed online 15 February, 2021).

Table 5
Description of the Emergent dataset: number of documents and distribution of assigned labels.

	News Bodies	Headlines	Claims	For	Against	Observing	Total Examples
Train	2,048	2,023	240	992	304	775	2,071
Test	523	513	60	246	91	187	524
Complete dataset	2,571	2,536	300	1,238	395	962	2,595

Table 6
Stance Stage results: class-wise F_1 Score, F_1m and overall accuracy on FNC-1 and Emergent dataset.

Experiment	F_1 Score			F_1m	Acc
	Agree	Disagree	Discuss		
Testing Emergent					
Emergent Upper Bound	81.53	74.53	68.23	74.76	76.15
Stance Stage Emergent	75.15	77.77	65.49	72.80	71.89
Stance Stage Emergent Test FNC-1 Training	73.15	73.68	70.61	72.48	72.08
Testing FNC-1					
Stance Stage FNC-1	72.87	63.50	88.74	75.04	82.30

Table 7
HeadlineStanceChecker results and comparison performance for the FNC-1 dataset.

System	F_1 Score				F_1m	Acc.	Rel. Score
	Agree	Disagree	Discuss	Unrelated			
Talos [45]	53.90	3.54	76.00	99.40	58.21	89.08	82.02
Athene [46]	48.70	15.12	78.00	99.60	60.40	89.48	82.00
UCLMR [47]	47.94	11.44	74.70	98.90	58.30	88.46	81.72
Human Upper Bound [46]	58.80	66.70	76.50	99.70	75.40	–	85.90
Dulhanty et al. [49]	**73.76**	55.26	85.53	99.12	78.42	93.71	90.00
Zhang et al. [48]	67.47	**81.30**	83.90	**99.73**	**83.10**	93.77	89.30
HeadlineStanceChecker-1stage	70.34	53.42	85.30	99.41	77.12	93.64	89.80
HeadlineStanceChecker-2stages	72.34	62.53	**87.32**	99.38	80.39	**94.31**	**91.02**

were calculated using the confusion matrices and results were published [47] or made available by the authors.[13,14]

The fourth row corresponds to the *Human Upper Bound*, and is the result of conducting the FNC-1 stance detection task manually. This upper bound was defined by [46]. Five human annotators were asked to manually label 200 random instances, obtaining an overall inter-annotator agreement of Fleiss'k of 0.686. Due to the fact that there is no upper bound reported in the FNC-1 data, we also used these values as reference for comparison purposes.

Next, the fifth and sixth rows include the results of recent approaches [48,49] that also addressed the headline stance detection task using the FNC-1 dataset, but did not take part in the challenge. Since there was no public code available, these results were also calculated from the confusion matrices provided in the papers.

The seventh row indicates the results for our *HeadlineStanceChecker* approach but configured only with a single classifier. We have called this approach *HeadlineStanceChecker-1stage*. Finally, the last row belongs to our *HeadlineStanceChecker* approach, using our proposed two-stage classification. For clarity purposes, in the table we will refer to this approach as *HeadlineStanceChecker-2stages*. Regardless of whether the classification is conducted in 1 or 2 stages, both approaches have used just the automatic summaries created from the full body text during the whole process.

As can be seen in Table 7, *HeadlineStanceChecker-2stages* is competitive enough with respect to the other systems, given that it only uses short summaries for the classification process, and not the full body text as the other systems use, so the information reduction does not imply a high loss in the results obtained, being better than the FNC-1 participants, and the human upper bound. Furthermore, the results also validate the fact that the divide-and-conquer strategy applied for dividing the classification into two stages is beneficial and yields better performance when using our proposed model with a single classifier (rows 7th and 8th). This is especially the case for detecting disagreement between the headline and the news article.

Furthermore, the most remarkable improvement for *HeadlineStanceChecker-2stages* is achieved in the *discuss* category, over performing all the remaining approaches. The F_1 improves by around 2 points compared to the second-best approach, i.e., [49], and 13 points over the lowest-performance system [47] in this category. By achieving competitive values in the other classes as well, *HeadlineStanceChecker-2stages* obtains a final macro-F1 value of 80.39%, being only beaten by the system proposed in [48], which takes advantage of a considerable number of external features beyond similarity to enrich the neural model. It is worth highlighting that in terms of accuracy and relative score, our approach (i.e., *HeadlineStanceChecker-2stages*) obtains the best result among the automatic systems in both cases, achieving 94.31% and 91.02%, respectively.

Focusing on the results obtained by the participants in the FNC-1 competition, when these results are analyzed independently for each of the classes, it can be seen that except for the classification of *unrelated* headlines (whose results are close to 100% in F1 measure, and this happens also for the remaining approaches as well) for the remaining classes, the results are very limited. The systems that participated in the FNC-1 competition have a very reduced performance especially in detecting the *disagree* stance, whereas the detection of *agree* is around 50% in F1 measure and for *discuss* around 75% for the best approach.

13 https://github.com/hanselowski/athene_system/ (accessed online 15 February, 2021).

14 https://github.com/Cisco-Talos/fnc-1 (accessed online 15 February, 2021).

Table 8
Class distribution for FNC-1 *subset>512* and FNC-1 *subset<512*.

	FNC-1 *subset>512*				FNC-1 *subset<512*			
	Agree	Disagree	Discuss	Unrelated	Agree	Disagree	Discuss	Unrelated
Train set	1,112	314	3,536	12,886	2,566	526	5,373	23,659
Test set	645	321	1,259	5,501	1,258	376	3,205	12,848
Total	1,757	635	4,795	18,387	3,824	902	8,578	36,507

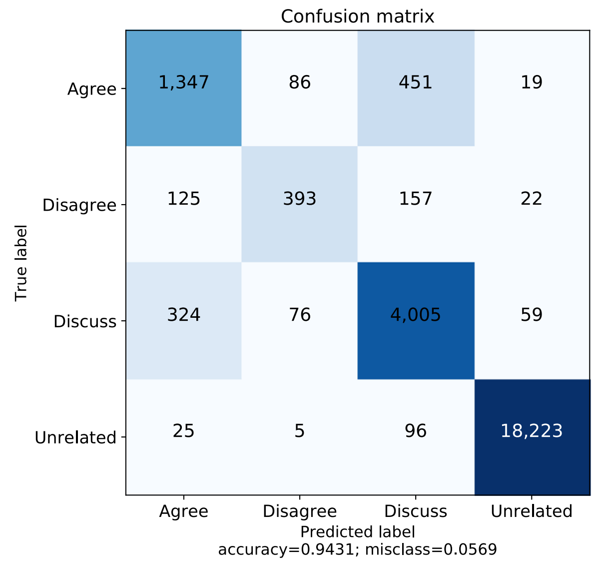

Fig. 5. Confusion matrix resulting from the *HeadlineStanceChecker-2stages*.

Table 9
HeadlineStanceChecker-2stages results for *subset>512* with different inputs: news body and news summary.

Input	F_1 Score				F_1m
	Agree	Disagree	Discuss	Unrelated	
News body	54.45	12.69	78.97	**99.52**	61.40
News summary	**59.61**	**28.06**	**80.85**	99.32	**66.96**

Outside the FNC-1 competition, the performance increases in all categories, being the disagree category one of the most challenging to classify, in which only the approach proposed in [48] obtains surprisingly high results for this category compared to the remaining methods.

After having established that *HeadlineStanceChecker-2stages* performs adequately (correctly detecting 94.31% of the test set classes), the confusion matrix presented in Fig. 5 provides more detail on the actual performance of the system for each stance class. From this information, we can observe that per class, the major classification problems occur with *disagree* and *discuss* classes. The data reflects that 22.5% of the total number of *disagree* stances are being classified as *discuss*, whereas as 23.6% of the total number of *agree* stances are classified as *discuss*. However, only 7.2% of the total number of *discuss* stances are being classified as *agree* and 17.9% of the total number of *disagree* stances are being classified as *agree*.

5.4. Summary versus body text analysis

Finally, in order to test the convenience of using the summary or the body text as input to our whole system, a final analysis was performed by an experiment designed to allow us to compare the results in both cases. To determine how this would be accomplished, we considered the singularities of our system, since the use of RoBERTa implies certain constraints that affect the input processing.

Table 10
HeadlineStanceChecker-2stages results for *subset<512* with different inputs: news body text and news summary.

Input	F_1 Score				F_1m
	Agree	Disagree	Discuss	Unrelated	
News body	**78.64**	**69.38**	**89.81**	**99.59**	**84.35**
News summary	74.17	58.91	87.69	99.36	80.03

RoBERTa, as a classification model, allows a maximum input length of 512 tokens, called *maximum sequence length*. Information that exceeds such a length is truncated. Our configuration takes as input both the headline and the text to which it refers, body text or summary, but it is relevant to remark that, in this case, the sequence length includes the tokens of the headline plus the tokens of that text. Since the headline must remain complete for the classification process, if it is necessary to truncate, it is the information in the body text which is lost.

In relation to the aforementioned issue, the previous work described in [49] focused on the analysis of the length of the body text for classification purposes, showing that for the examples in which the input sequence is greater than 512 tokens, the accuracy of the classification decreases considerably with respect to smaller sequences.

Against this backdrop, our hypothesis states that applying summarization to the text before classification implies an improvement in the results. In order to prove it, we first create two subsets from the FNC-1 dataset according to the news story length: *subset>512* and *subset<512*. Table 8 shows the class distribution for both subsets.

Next, we trained and tested the system with both subsets twice: first with the bodies as input, and then with the summaries. The results in Table 9 show that for long news stories (*subset>512*), the system performs better with summaries as input than truncating the text of the full article. This could happen because reducing the input by simply cutting text at the end of the document results in relevant information being lost, whereas when making a summary, it is the relevant information that prevails in a more concise mode.

Similarly, results for *subset<512*, the shortest news stories, are reported in Table 10. The system was again trained and tested, taking the body and the summary as inputs. In this case, results are better when using the full body text, which could indicate that all the information needed for a proper classification is present by considering the whole text (an unfeasible scenario with longer texts).

There exist no explicit rules that determine what the length of a news article should be, but there is instead certain evidence supporting that news tend to be longer than 512 tokens. In Table 11 we have gathered statistics from the most popular news datasets that are being used in language processing tasks. All together, they contain more than 2 million articles from different sources, with an average length superior to 512 tokens. The relevance of our approach is made clear by these figures, which indicate that, in most cases, using news summarization would be the right strategy.

Table 11
Statistics from large news corpora indicating the average document length in words.

	Examples	Average length
CNN [103]	92 K	760.50
DailyMail [103]	310 K	653.33
NY times [104]	650 K	800.04
Newsroom [105]	1,210 K	770.09
Total	2,260 K	745.83

6. Conclusions and future work

HeadlineStanceChecker has been demonstrated to be an effective approach for detecting misinformation in news, specifically when a headline has to be compared to its body text. The novelty of our approach rests on two key premises: (i) the adoption of a divide-and-conquer strategy, thus tackling the stance classification problem by means of a 2-stage neural architecture; and (ii) the use of extractive semantic summarization instead of the full news body text for the whole classification, in addition to a salience and two similarity features that will help to determine the relatedness of the headline with respect to the news article.

To show the appropriateness of *HeadlineStanceChecker*, different experiments were carried out in the context of an existing task (Fake News Challenge FNC-1), where the stance of a headline had to be classified into one of the following classes: *unrelated*, *agree*, *disagree*, and *discuss*. The experiments involved validating each of the proposed classification stages in isolation together with the whole approach, as well as a comparison with respect the state of the art in this task. Furthermore, additional experiments with another corpus for headline stance detection (i.e., Emergent dataset) were also performed to verify the generalization of our approach. The results obtained by our system were very competitive compared to other SOTA systems obtaining 94.31% Accuracy, as well as the highest result in FNC-1 relative score compared with the state of the art (91.02%).

The unbalanced nature of the FNC-1 dataset leads to existing systems being more capable of learning how to detect *unrelated* headlines, but are less accurate when it comes to the remaining classes. Even so, the results obtained by *HeadlineStanceChecker* for the different categories with less examples, *agree, disagree* and *discuss* are fair enough and promising, which indicates that the chosen approach is appropriate for the task.

Future work will aim to improve the results of *agree* and *disagree* classification by extending the system to take into consideration Sentiment Analysis features. Furthermore, as reported speech is recently being introduced to determine bias and document stance, it could be very useful for determining the stance of headlines and news articles. Some reporting events are neutral, for example, by using *reported* or *said*, whereas some others introduce a stance, for instance, *'deny'* implies disagreement or *'confirm'* indicates agreement.

Besides, another interesting aspect to focus on would be to investigate the relation of the stance detection classes (agree, disagree, discuss and unrelated) with the "incongruent" and "congruent" classification to determine whether this relation can provide some insights for different scenarios.

Finally, as a future goal that contributes to investigating the problem of fake news detection, we expect to apply *HeadlineStanceChecker* to a real world scenario to detect when headlines introduce mis- or disinformation to readers. Our contribution to improving the current research in the field, by means of new learning strategies and discourse aware techniques, will help to combat online fake news, a societal problem that requires concerted action.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research work has been partially funded by Generalitat Valenciana through project "SIIA: Tecnologías del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible" with grant reference PROMETEU/2018/089, by the Spanish Government through project RTI2018-094653-B-C22: "Modelang: Modeling the behavior of digital entities by Human Language Technologies", as well as by the project RTI2018-094649-B-I00: "INTEGER - Intelligent Text Generation". Also, this paper is also based upon work from COST Action CA18231 "Multi3Generation: Multi-task, Multilingual, Multi-modal Language Generation".

References

[1] V.L. Rubin, Disinformation and misinformation triangle, J. Doc. 75 (5) (2019) 1013–1034.
[2] M. Tudjmanand, N. Mikelic Preradovic, Information Science: Science about Information, in: Proceedings of Informing Science & IT Education, 2003, pp. 1513–1527.
[3] S. Tavernisen, As fake news spreads lies, more readers shrug at the truth, N.Y. Times (2019).
[4] A. Bovet, H.A. Makse, Influence of fake news in Twitter during the 2016 US presidential election, Nature Commun. 10(1):7 (2019).
[5] M.T. Bastos, D. Mercea, The brexit botnet and user-generated hyperpartisan news, Soc. Sci. Comput. Rev. 37 (1) (2019) 38–54.
[6] V. Hooper, Fake news and social media: The role of the receiver, in: 5th European Conference on Social Media 2018, Academic Conferences and publishing limited, 2018, p. 62.
[7] S. Issenberg, The Victory Lab: The Secret Science of Winning Campaigns, Crown, 2012.
[8] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, M. Palomar, Fighting post-truth using natural language processing: a review and open challenges, Expert Syst. Appl. 141 (2020) 112943.
[9] T. van Dijk, News As Discourse, in: Communication Series, L. Erlbaum Associates, 1988.
[10] J. Kuiken, A. Schuth, M. Spitters, M. Marx, Effective headlines of newspaper articles in a digital environment, Digit. J. 5 (10) (2017) 1300–1314.
[11] M. Gabielkov, A. Ramachandran, A. Chaintreau, A. Legout, Social clicks: What and who gets read on Twitter? ACM SIGMETRICS Perform. Eval. Rev. 44 (2016) 179–192.
[12] B. Lutz, M.T. Adam, S. Feuerriegel, N. Pröllochs, D. Neumann, Affective information processing of fake news: Evidence from neurois, in: Information Systems and Neuroscience, Springer International Publishing, 2020, pp. 121–128.
[13] Y. Chen, N.J. Conroy, V.L. Rubin, News in an online world: The need for an "automatic crap detector", in: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, American Society for Information Science, 2015.
[14] W. Wei, X. Wan, Learning to identify ambiguous and misleading news headlines, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 4172–4178.
[15] Y. Chen, N.J. Conroy, V.L. Rubin, Misleading online content: Recognizing clickbait as "false news", in: Proceedings of the ACM on Workshop on Multimodal Deception Detection, Association for Computational Linguistics, 2015, pp. 15–19.
[16] S. Chesney, M. Liakata, M. Poesio, M. Purver, Incongruent Headlines: Yet Another Way to Mislead Your Readers, in: Proceedings of Natural Language Processing Meets Journalism, 2017, pp. 56–61.
[17] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2016, pp. 1163–1168.
[18] Y. Hayashi, H. Yanagimoto, Headline generation with recurrent neural network, in: New Trends in E-Service and Smart Computing, Springer, 2018, pp. 81–96.
[19] Z. Huang, Z. Ye, S. Li, R. Pan, Length adaptive recurrent model for text classification, in: Proceedings of the ACM on Conference on Information and Knowledge Management, Association for Computing Machinery, 2017, pp. 1019–1027.

[20] M. Choraś, K. Demestichas, A. Giełczyk, A. Herrero, P. Ksieniewicz, K. Remoundou, D. Urda, M. Woźniak, Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study, Appl. Soft Comput. (2020) 107050.

[21] G. Di Domenico, J. Sit, A. Ishizaka, D. Nunan, Fake news, social media and marketing: A systematic review, J. Bus. Res. 124 (2021) 329–341.

[22] G. Zarrella, A. Marsh, Mitre at SemEval-2016 task 6: Transfer learning for stance detection, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, 2016, pp. 458–463.

[23] S. Ghosh, P. Singhania, S. Singh, K. Rudra, S. Ghosh, Stance detection in web and social media: a comparative study, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 75–87.

[24] S. Somasundaran, J. Wiebe, Recognizing stances in ideological on-line debates, in: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches To Analysis and Generation of Emotion in Text, 2010, pp. 116–124.

[25] A. Konjengbam, S. Ghosh, N. Kumar, M. Singh, Debate stance classification using word embeddings, in: International Conference on Big Data Analytics and Knowledge Discovery, Springer, 2018, pp. 382–395.

[26] A. Faulkner, Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure, in: The Twenty-Seventh International Flairs Conference, 2014.

[27] C. Li, A. Porco, D. Goldwasser, Structured representation learning for online debate stance prediction, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3728–3739.

[28] R. Agrawal, S. Rajagopalan, R. Srikant, Y. Xu, Mining newsgroups using networks arising from social behavior, in: Proceedings of the 12th International Conference on World Wide Web, 2003, pp. 529–535.

[29] A. Murakami, R. Raymond, Support or oppose? classifying positions in online debates from reply activities and opinion expressions, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 869–875.

[30] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 845–854.

[31] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, Semeval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEVal-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 31–41.

[32] A.I. Al-Ghadir, A.M. Azmi, A. Hussain, A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments, Inf. Fusion 67 (2021) 29–40.

[33] B.G. Patra, D. Das, S. Bandyopadhyay, JU_NLP at SemEval-2016 task 6: detecting stance in tweets using support vector machines, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 440–444.

[34] H. Elfardy, M. Diab, Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 434–439.

[35] I. Augenstein, T. Rocktäschel, A. Vlachos, K. Bontcheva, Stance detection with bidirectional conditional encoding, 2016, arXiv preprint arXiv:1606.05464.

[36] P. Wei, W. Mao, D. Zeng, A target-guided neural memory model for stance detection in Twitter, in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

[37] S. Zhou, J. Lin, L. Tan, X. Liu, Condensed convolution neural network by attention over self-attention for stance detection in twitter, in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[38] A. Sen, M. Sinha, S. Mannarswamy, S. Roy, Stance classification of multi-perspective consumer health information, in: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, 2018, pp. 273–281.

[39] M. Taulé, M.A. Martí, F.M. Rangel, P. Rosso, C. Bosco, V. Patti, et al., Overview of the task on stance and gender detection in tweets on catalan independence at IberEval 2017, in: 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, 1881, CEUR-WS, 2017, pp. 157–177.

[40] S.V. Vychegzhanin, E.V. Kotelnikov, Stance detection based on ensembles of classifiers, Program. Comput. Softw. 45 (5) (2019) 228–240.

[41] M. Lai, A.T. Cignarella, D.I.H. Farías, C. Bosco, V. Patti, P. Rosso, Multilingual stance detection in social media political debates, Comput. Speech Lang. 63 (2020) 101075.

[42] E. Zotova, R. Agerri, G. Rigau, Semi-automatic generation of multilingual datasets for stance detection in Twitter, Expert Syst. Appl. 170 (2021) 114547.

[43] M. Babakar, N. Bakos, H. Daumé, A. Mantzarlis, D. Seddah, A. Vlachos, C. Wardle, Fake news challenge - I, 2016.

[44] C. Silverman, Lies, damn Lies and viral content, 2019.

[45] S. Baird, D. Sibley, Y. Pan, Talos targets disinformation with fake news challenge victory, 2017, https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html, last accessed on 20/02/21.

[46] B.S. Andreas Hanselowski, F. Caspelherr, Description of the system developed by team athene in the FNC-1, 2017.

[47] B. Riedel, I. Augenstein, G.P. Spithourakis, S. Riedel, A simple but tough-to-beat baseline for the fake news challenge stance detection task, Comput. Res. Repos., CoRR (2017) arXiv:1707.03264.

[48] Q. Zhang, S. Liang, A. Lipani, Z. Ren, E. Yilmaz, From stances' imbalance to their hierarchical representation and detection, in: The World Wide Web Conference, ACM, 2019, pp. 2323–2332.

[49] C. Dulhanty, J.L. Deglint, I.B. Daya, A. Wong, Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection, 2019, arXiv preprint arXiv:1911.11951.

[50] B. Pouliquen, R. Steinberger, C. Best, Automatic detection of quotations in multilingual news, 2007, pp. 487–492.

[51] A. Vlachos, S. Riedel, Identification and Verification of Simple Claims about Statistical Properties, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2596–2601.

[52] M.-C. De Marneffe, A.N. Rafferty, C.D. Manning, Finding contradictions in text, Proc. Assoc. Comput. Linguist. (2008) 1039–1047.

[53] S. Harabagiu, A. Hickl, F. Lacatusu, Negation, contrast and contradiction in text processing, in: AAAI, 6, 2006, pp. 755–762.

[54] O. Levy, T. Zesch, I. Dagan, I. Gurevych, Recognizing partial textual entailment, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 2, 2013, pp. 451–455.

[55] S.A. Brown, The effects of explicit main idea and summarization instruction on reading comprehension of expository text for alternative high school students, Ph.D. thesis, Utah State University, 2018.

[56] J. Engelen, G. Camp, J. van de Pol, A. de Bruin, Teachers' monitoring of students' text comprehension: can students' keywords and summaries improve teachers' judgment accuracy? Metacognition Learn. 13 (3) (2018) 287–307.

[57] X. Lin, S.-E. Jhang, D. Dong, Investigating the effects of text summarization on linguistic quality of argumentative writing, 60, (4) 2018, pp. 245–268,

[58] J.P. Barreiro, Improving reading comprehension of narrative texts through summaries, Ph.D. thesis, Universidad Casa Grande, 2019.

[59] R. Dijkman, A. Wilbik, Linguistic summarization of event logs – a practical approach, Inf. Syst. 67 (2017) 114–125.

[60] J. Petkovic, V. Welch, M. Jacob, M. Yoganathan, A.P. Ayala, H. Cunningham, P. Tugwell, The effectiveness of evidence summaries on health policymakers and health system managers use of evidence from systematic reviews: A systematic review, Implement. Sci. 11 (2016).

[61] L. Hartling, A. Gates, J. Pillay, M. Nuspl, A. Newton, Development and usability testing of EPC evidence review dissemination summaries for health systems decisionmakers., methods research report, Technical Report., 2018.

[62] Y. Liu, X. Song, S.-F. Chen, Long story short: finding health advice with informative summaries on health social media, Aslib J. Inf. Manag. ahead-of-print (2019).

[63] F. Dernoncourt, M. Ghassemi, W. Chang, A repository of corpora for summarization, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[64] A. Nenkova, Automatic text summarization of newswire: Lessons learned from the document understanding conference, in: Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3, in: AAAI'05, AAAI Press, 2005, pp. 1436–1441.

[65] S. Mackie, R. McCreadie, C. Macdonald, I. Ounis, Experiments in newswire summarisation, in: N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G.M. Di Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2016, pp. 421–435.

[66] Y. Duan, A. Jatowt, Across-time comparative summarization of news articles, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, in: WSDM, Association for Computing Machinery, 2019, pp. 735–743.

[67] C. Zhu, Z. Yang, R. Gmyr, M. Zeng, X. Huang, Make lead bias in your favor: A simple and effective method for news summarization, 2019, arXiv preprint arXiv:1912.11602.

[68] M. Banko, V.O. Mittal, M.J. Witbrock, Headline generation based on statistical translation, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2000, pp. 318–325.

[69] B. Dorr, D. Zajic, R. Schwartz, Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation, in: Proceedings of the North American Chapter of the Association for Computational Linguistics, Text Summarization Workshop, 2003, pp. 1–8.
[70] D. Zajic, B. Dorr, R. Schwartz, Automatic Headline Generation for Newspaper Stories, in: Proceedings of the Workshop on Automatic Summarization, 2002, pp. 78–85.
[71] J. Tan, X. Wan, J. Xiao, From neural sentence summarization to headline generation: A coarse-to-fine approach, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 4109—4115.
[72] D. Gavrilov, P. Kalaidin, V. Malykh, Self-attentive model for headline generation, in: Advances in Information Retrieval, Springer International Publishing, 2019, pp. 87–93.
[73] K. Iwama, Y. Kano, Multiple news headlines generation using page metadata, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, 2019, pp. 101–105.
[74] S. Esmaeilzadeh, G.X. Peh, A. Xu, Neural abstractive text summarization and fake news detection, Comput. Res. Repos. CoRR (2019) arXiv:1904.00788.
[75] G. Kim, Y. Ko, Graph-based Fake News Detection using a Summarization Technique, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 3276–3280.
[76] P. Krejzl, B. Hourová, J. Steinberger, Stance detection in online discussions, Comput. Res. Repos. CoRR (2017) arXiv:1701.00504.
[77] P. Krejzl, Stance detection and summarization in social networks, Report, Západočeská univerzita v Plzni, 2018.
[78] S. Chesney, M. Liakata, M. Poesio, M. Purver, Incongruent headlines: Yet another way to mislead your readers, in: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 56–61, http://dx.doi.org/10.18653/v1/W17-4210.
[79] D. Küçük, F. Can, Stance detection: A survey, ACM Comput. Surv. 53 (1) (2020) 1–37.
[80] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A survey on stance detection for mis-and disinformation identification, 2021, arXiv preprint arXiv:2103.00242.
[81] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1163–1168, http://dx.doi.org/10.18653/v1/N16-1138.
[82] Y. Lv, C. Zhai, Positional language models for information retrieval, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 299–306.
[83] M. Vicente, C. Barros, E. Lloret, Statistical language modelling for automatic story generation, J. Intell. Fuzzy Systems 34 (5) (2018) 3069–3079.
[84] A. Kilgarriff, C. Fellbaum, WordNet: An electronic lexical database, Language 76 (3) (2000) 706, http://dx.doi.org/10.2307/417141.
[85] M.E. Vicente, E. Lloret, A discourse-informed approach for cost-effective extractive summarization, in: L.E. Anke, C. Martín-Vide, I. Spasic (Eds.), Statistical Language and Speech Processing - 8th International Conference, Proceedings, in: Lecture Notes in Computer Science, 12379, Springer, 2020, pp. 109–121.
[86] M. Vicente, R. Sepúlveda-Torrres, C. Barros, E. Saquete, E. Lloret, Can text summarization enhance the headline stance detection task? benefits and drawbacks, in: International Conference on Document Analysis and Recognition, Springer International Publishing, 2021, pp. 53–67.
[87] L. Padró, E. Stanilovsky, Freeling 3.0: Towards wider multilinguality, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey, 2012.
[88] N. Passalis, A. Tefas, Learning bag-of-embedded-words representations for textual information retrieval, Pattern Recognit. 81 (2018) 254–267.
[89] B. Li, L. Han, Distance weighted cosine similarity measure for text classification, in: Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning, Springer-Verlag, 2013, pp. 611–618.
[90] S. Tata, J.M. Patel, Estimating the selectivity of tf-idf based cosine similarity predicates, SIGMOD Rec. 36 (4) (2007) 75–80.
[91] V. Kotu, B. Deshpande, Classification, in: Data Science, Elsevier, 2019, pp. 65–163.
[92] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.
[93] F. Šarić, G. Glavaš, M. Karan, J. Šnajder, B. Dalbelo Bašić, Takelab: Systems for measuring semantic text similarity, in: Proceedings of the First Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, 2012, pp. 441–448.
[94] L. Metcalf, W. Casey, Metrics, similarity, and sets, in: Cybersecurity and Applied Mathematics, Elsevier, 2016, pp. 3–22.
[95] M. Vijaymeena, K. Kavitha, A survey on similarity measures in text mining, Mach. Learn. Appl. Int. J. 3 (2) (2016) 19–28.
[96] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[97] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, 2018, pp. 353–355.
[98] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 785–794.
[99] V. Slovikovskaya, Transfer learning from transformers to fake news challenge stance detection (FNC-1) task, 2019, arXiv preprint arXiv:1910.14353.
[100] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[101] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, 2019, ArXiv arXiv:1910.03771.
[102] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C.M. Meyer, I. Gurevych, A retrospective analysis of the fake news challenge stance-detection task, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, 2018, pp. 1859–1874.
[103] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Advances in Neural Information Processing Systems, 2015, pp. 1693–1701.
[104] E. Sandhaus, The new york times annotated corpus LDC2008t19, Linguistic Data Consortium, Philadelphia, 2008.
[105] M. Grusky, M. Naaman, Y. Artzi, Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, 2018, arXiv preprint arXiv:1804.11283.