Entity-centric Reasoning and Claim Validation based on Dynamic Textually Represented Environments

Dimitris Papadopoulos
Technical University of Crete
Chania, Greece
dpapadopoulos6@isc.tuc.gr

Katerina Metropoulou
National Technical University of Athens
Athens, Greece
kmetropoulou@mail.ntua.gr

Nikolaos Matsatsinis
Technical University of Crete
Chania, Greece
nmatsatsinis@isc.tuc.gr

Nikolaos Papadakis
Hellenic Army Academy
Vari, Greece
npapadakis@sse.gr

Abstract

Our collective attention span is shortened by the flood of online information. We address the need for automated claim validation based on the unstructured, aggregated information, continuously derived from multiple online news sources. We introduce an entity-centric reasoning framework in which latent connections between events, actions, or statements are revealed via entity mentions and are structurally integrated in a graph database. Using entity linking and semantic similarity, we collect and combine information from diverse sources, in order to generate evidence relevant to the user's claim. We then leverage textual entailment recognition to quantitatively determine whether this assertion is credible, based on the created evidence. Our approach fills the gap in automated claim validation for less-resourced languages, while being showcased for the Greek language and is complemented by the training of semantic textual similarity (STS) and natural language inference (NLI) models which are evaluated on translated versions of common benchmarks.

1 Introduction

Motivation: The wider diffusion of the Web since the dawn of Web 2.0 has enabled instantaneous access to an expanding universe of information. The entire nature of news consumption has shifted dramatically, as individuals increasingly rely on the Internet as their major source of information. While people access, filter and blend several websites into intricate patterns of media consumption, this wealth of unstructured information contained in billions of online articles inevitably creates a poverty of attention and a need to efficiently allocate this attention among the many sources that may absorb it. Verifying whether a given claim coheres with the knowledge hidden in the vast amount of published information is a fundamental problem in NLP, especially when taking into account that the arrival of new information may weaken or retract the initially supported inference.

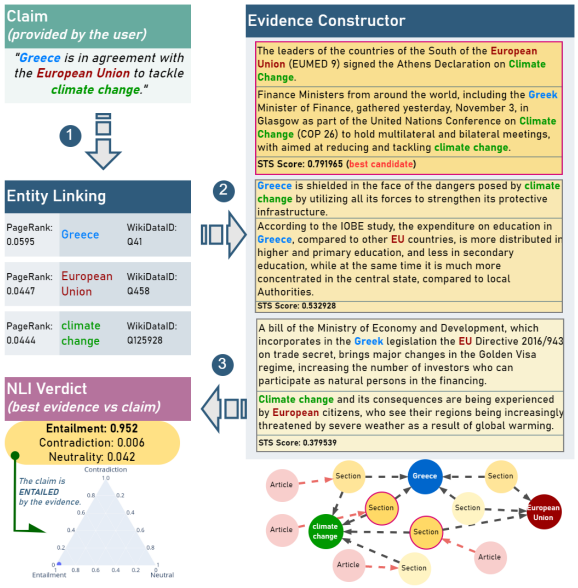

Figure 1: Claim validation example (translated from Greek) based on aggregated evidence using *FarFetched*.

Approach and Contribution: *FarFetched* is a modular framework that enables people to verify any kind of textual claim based on the incorporated evidence from textual news sources. It combines a series of processes to periodically crawl for news articles and annotate their context with named entities. Given a user claim, *FarFetched* derives a relevant subset of the stored content based on its semantic similarity with the provided claim, thus being able to reason about its validity in an NLI setting (Figure 1). While the proposed framework

focuses on the less-resourced Greek language, its modular architecture allows the integration of pre-trained models for any language. Moreover, it is capable of topic-agnostic, evidence-aware assessment of arbitrary textual claims in a fully automated manner, without relying on feature engineering, curated sources and manual intervention.

The main contributions of this work are summarized as follows: a) to formalize, develop and evaluate a claim validation and reasoning approach based on the aggregated knowledge derived from the continuous monitoring of news sources, and b) to train, evaluate and share SotA models for the STS and NLI downstream tasks for the Greek language that support the core functionalities of our framework.[1]

2 Related Work

Our work comprises functionalities comparable to those of fact checking frameworks, targeting the assignment of a truth value to a claim made in a particular context (Vlachos and Riedel, 2014). For most related approaches (Zhang et al., 2021; Majithia et al., 2019; Zhou et al., 2019; Ciampaglia et al., 2015; Goasdoué et al., 2013) the evidence to support or refute a claim is derived from a trustworthy source (e.g. Wikipedia, crowdsourced tagging or expert annotators). Interesting deviations are DeClarE (Popat et al., 2018) that searches for web articles related to a claim considering their in-between relevance using an attention mechanism, and ClaimEval (Samadi et al., 2016), based on first-order logic to contextualise prior knowledge from a set of the highest page-ranked websites.

FarFetched can be distinguished from the aforementioned works by four major points: a) evidence collection is disentangled from manual annotation but relies on a constantly updating feed of news articles instead; b) claim validation based on the accumulated evidence relies on the effective combination of entity linking and attention-based models; c) our approach provides interpretable reasoning based on the aggregated evidence of multiple sources without assessing their truthfulness as opposed to most fact checking frameworks; and d) the outcome of the process is dynamic as the continuous integration of new information may lead to a shift in the verdict of the validated claim.

Recent advances in the field of *event-centric* NLP have introduced event representation methods based on narrative event chains (Vossen et al., 2015), knowledge graphs (Tang et al., 2019; Vossen et al., 2016), QA pairs (Michael et al., 2018) or event network embeddings (Zeng et al., 2021) to capture connections among events in a global context. Our method relies on an *entity-centric* approach instead, where the identified entities are used as connectors between events, actions, facts, statements or opinions, thus revealing latent connections between the articles containing them. A few similar approaches have been proposed for combining world knowledge with event extraction methods to represent coherent events, but rely either on causal reasoning to generate plausible predictions (Radinsky et al., 2012) or on QA models that require the accompanying news source to be provided along with the user's question (Jin et al., 2021).

The latest advances regarding the technological concepts that comprise our methodology are provided below:

Entity linking (EL) resolves the lexical ambiguity of entity mentions and determines their meanings in context. Typical EL approaches aim at identifying named entities in mention spans and linking them to entries of a KG (e.g. Wikidata, DBpedia) thus resolving their ambiguity. Recent methods combine the aforementioned tasks using local compatibility and topic similarity features (Delpeuch, 2019), pagerank-based wikification (Brank et al., 2017a) —used also in *FarFetched*— or neural end-to-end models that jointly detect and disambiguate mentions with the help of context-aware mention embeddings (Kolitsas et al., 2018).

The recent interest for encapsulating diverse semantic sentence features into fixed-size vectors has resulted in SotA systems for Semantic Textual Similarity (STS) based on supervised cross-sentence attention (Raffel et al., 2020), Deep Averaging Networks (DAN) (Cer et al., 2018) or siamese and triplet BERT-Networks (Reimers and Gurevych, 2019) to acquire meaningful sentence embeddings that can be compared using cosine similarity. The latter approach is leveraged in our case to train an STS model for the Greek language using transfer learning.

Finally, the task of Natural Language Inference (NLI) -also known as Recognizing Textual Entailment (RTE)- associates an input pair of premise and hypothesis phrases into one of three classes:

[1]Code and translated benchmark datasets: https://github.com/lighteternal/FarFetched_NLP

contradiction, entailment and neutral. Ferreira and Vlachos, 2016 modeled fact checking as a form of RTE to predict whether a premise, typically part of a trusted source, is for, against, or observing a given claim. SotA NLI models typically rely on Transformer variants with global attention mechanisms (Beltagy et al., 2020), siamese network architectures (Reimers and Gurevych, 2019) (also used in *FarFetched* to train a Greek NLI model), autoregressive language models for capturing long-term dependencies (Yang et al., 2019) and denoising autoencoders (Lewis et al., 2020).

3 Methodology

3.1 Problem Definition

Given a user claim in free text, we tackle the problem of deciding whether this statement is plausible based on the currently accumulated knowledge from news sources. We also acknowledge the problem of constructing relevant evidence from multiple sources by analysing the information contained in online articles and the need for efficiently extracting only contextually and semantically relevant excerpts to verify or refute the user's claim. While our work does not primarily focus on better sentence embeddings and natural language inference techniques, we also target the lack of such models for the Greek language.

3.2 Our approach

FarFetched combines a series of *offline* (i.e. performed periodically) operations to accumulate data from various news sources and annotate their context with named entities. It also encompasses a number of *online* operations (i.e. upon user input) to assess the validity of a claim in free text. First, it identifies the entities included in the provided claim and leverages these as a starting point to derive a relevant subset of the stored textual information as candidate evidence. Each candidate is then compared with the claim in terms of textual similarity, in order to finally conclude on the most relevant evidence (premise) to reason about the validity of the claim (hypothesis) in an NLI setting. The distinct modules that comprise the framework are visualised in Figure 2. The process that *FarFetched* follows to evaluate a claim is summarized in Algorithm 1, while each module is described in greater detail in the following subsections.

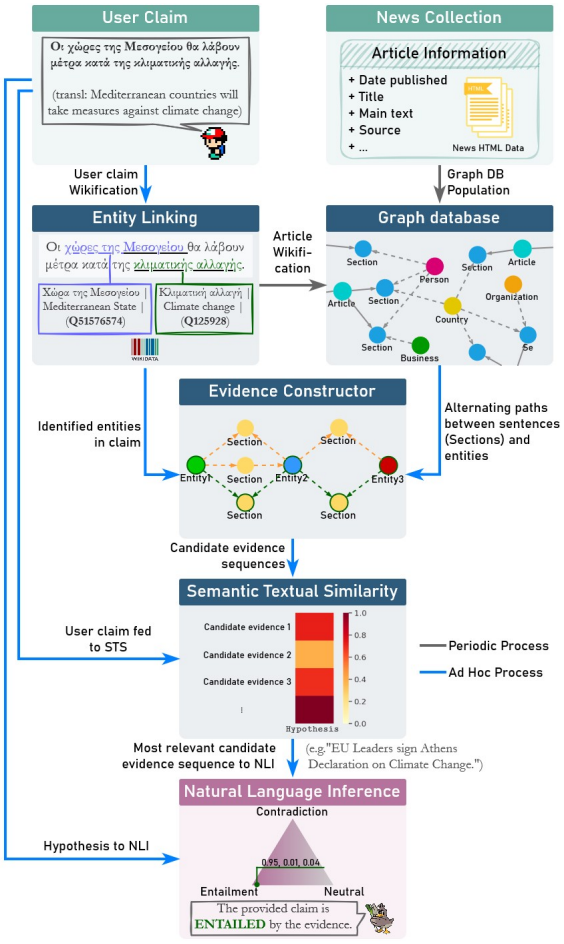

Figure 2: The *FarFetched* modular framework.

3.2.1 News Collection

A multilingual, open-source crawler and extractor for heterogeneous website structures is leveraged to incorporate information from various news sources (Hamborg et al., 2017). It is capable of extracting the major properties of news articles (i.e., title, lead paragraph, main content, publication date, author, etc.), featuring full website extraction and requiring only the root URL of a news website to crawl it completely.

3.2.2 Graph Database Population

The crawled articles are forwarded to a graph database (Webber, 2012) that initially stores only two types of nodes: `Article`, which represents a news article with its aforementioned properties and `Section` that represents a sentence of each article's main text (i.e. concatenated title and article body). Each `Article` node is linked to one or more `Section` nodes via the `HAS_SECTION` relationship.

Algorithm 1 Claim Evaluation

Input: A claim c provided by the user in natural language.

Output: Most relevant evidence seq^* (sequence of article excerpts) based on the input claim c along with its STS_score^* and $NLI\_score < c, e, n >$.

Entity Linking: Find the set of entities $(e_1, ..., e_n) \in E$, where $e \exists c$ and $|E| = n$
$S \leftarrow \varnothing$
Graph database search: Find all shortest paths p between the alternating entities e and sentences s: $p \leftarrow (e_1, s_a, e_2, ..., e_{n-1}, s_k, e_n) \in P$
if $P = \varnothing$ **then**
    $s \in P \iff s$ has at least 1 entity mention
end if
for $p_i \in P$ **do**
    $seq_i \leftarrow (s_a, ..., s_k)$
    $S \leftarrow S \cup seq_i$ (sequence seq_i added to candidate evidence set)
end for
$STS\_Scores \leftarrow \varnothing$
for $seq_i \in S$ **do**
    Semantic Textual Similarity: Compare $seq_i \in S$ to c (each candidate evidence sequence to the claim) and calculate STS_score_i
    $STS\_Scores \leftarrow STS\_Scores \cup STS\_score_i$
end for
Find the candidate seq^* with the highest similarity to the claim: $STS\_score^* \leftarrow max(STS\_Scores)$
$seq^* \leftarrow argmax(STS\_score^*)$
Natural Language Inference: Compare seq^* to c (the best candidate evidence to the claim) and calculate the scores for contradiction, entailment and neutrality $NLI\_score < c, e, n >$

3.2.3 Entity Linking

Given that our approach relies on largely unstructured textual documents that lack explicit semantic information, Entity Linking (EL) constitutes a central role in revealing latent connections between seemingly uncorrelated article sections. To this end, *FarFetched* employs a type of semantic enrichment and entity disambiguation technique known as wikification (Brank et al., 2017b), which involves using Wikipedia concepts as a source of semantic annotation. It applies pagerank-based wikification on input text to identify phrases that refer to entities of the target knowledge base (Wikipedia) and return their corresponding WikiData Entity ID. The latter is used as a unique identifier for storing the entities as `Entity` nodes to the graph database and for linking them with the crawled article `Section` nodes, resulting to a more tightly connected graph, where article sections are connected to WikiData entities via the `HAS_ENTITY` relationship. The virtual graph of Figure 3 represents the structure (labels and relationships) of the graph database. It should be noted that an entity node might have an additional label (e.g. `Person`, `City`, `Business`) except for the generic `Entity` one, based on the WikiData class taxonomy.

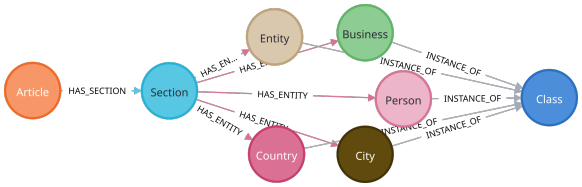

Figure 3: Final structure of the graph database.

3.2.4 Evidence Constructor

In a typical NLI setting, a premise represents our knowledge or evidence regarding an event and is used to infer whether a relevant hypothesis follows from it or not. In our case, article sections focusing on the same entities as the user's claim could potentially lead to the construction of useful evidence towards the validation of this claim. We can therefore leverage the entity-annotated article sections of our graph database to collect relevant evidence by aggregating information from multiple sources. To this end, we developed an evidence construction process that comprises the following steps:

1. The claim provided by the user passes through the Entity Linking phase and one or more entities (WikiData concepts) are identified.

2. The graph database is queried for all possible shortest paths that contain article sections between the identified entities. Given the implemented graph structure and n Entity nodes, this translates to a minimum path length of $2(n-1)$ alternating Entity-Section nodes as shown in Figure 4. Since the existence of such path is not guaranteed, in cases that no path is found the algorithm will select an article section if it contains at least one mentioned entity.

3. The article sections contained in these paths are concatenated to form a set of candidate evidence sequences. Their relevance with the claim at hand is assessed during the Semantic Textual Similarity phase.

3.2.5 Semantic Textual Similarity

We train and apply a sentence embeddings method to extract and compare the vector representations

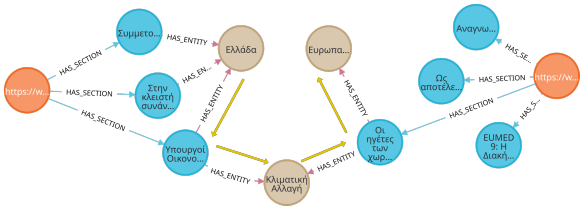

Figure 4: Shortest path example: The 3 entities (in brown) are connected with 2 article sections (in blue).

of the user's claim with each candidate evidence sequence, in order to select the most semantically relevant candidate for the final NLI phase. Despite the abundance of multilingual language models (e.g. m-BERT, XLM) that cover most common languages, pretrained multilingual sentence embeddings models do not generally perform well in downstream tasks for less-resourced languages like Greek (Koutsikakis et al., 2020). Furthermore, given that the vector spaces between languages are not aligned, sentences with the same content in different languages could be mapped to different locations in the common vector space. To overcome this obstacle, we trained a Greek sentence embeddings model on parallel EN-EL (English-Greek) sentence pairs following a multilingual knowledge distillation approach (Reimers and Gurevych, 2020a). Our Greek student model (XLM-RoBERTa) was trained on parallel pairs to produce vectors for the EN-EL sentences that are close to the teacher's pretrained English model ones (DistilRoBERTa). Using the trained model, we are able to compare the produced vector representations between the claim and each concatenated candidate evidence sequence with regard to STS in terms of cosine similarity and forward the best candidate to the last phase of the claim validation process, namely Natural Language Inference.

3.2.6 Natural Language Inference

The last step of our process leverages NLI to determine whether the user claim (hypothesis) is entailed by, contradicted, or neutral to the most relevant evidence (premise) of the previous phase. To tackle the aforementioned multilinguality issues of pretrained language models on less-resourced languages, we finetuned a Greek *sentence-transformers* Cross-Encoder (Reimers and Gurevych, 2019) (*XLM-RoBERTa-base*) model for the NLI task. The model was trained on the Greek and English version of the combined SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) corpora (AllNLI). We used the English-to-Greek machine translation model by Papadopoulos et al., 2021 to create the Greek version of the AllNLI dataset. The trained model takes the premise-hypothesis pair as input and predicts one of the following labels for each case: "contradiction": c, "entailment": e or "neutral": n. The logits for each class are then converted to probabilities using the softmax function. These labels along with their probability scores can be used to assess whether the claim is verified by the accumulated knowledge of the candidate evidence.

4 Experiments

4.1 Setup

The technical details for each building block of *FarFetched* are provided below:

News Collection and Storage: The *news-please* (Hamborg et al., 2017) Python library was used to ingest an initial corpus of news articles to support our experiments. The root URLs of two popular Greek news sites served as the starting point in order to recursively crawl news from a diverse topic spectrum, spanning from 2018 until 2021. We collected 13,236 articles, containing 31,358 sections in total. A *Neo4j* graph DBMS was used to store the crawled articles and sections as nodes and create their in-between relationships.

Entity Linking: A Python script producing POST requests to the *JSI Wikifier* web API (Brank et al., 2017a) was implemented to annotate the article sections and enrich the graph database with WikiData entities. A total of 2,516 WikiData entities of different types (e.g. sovereign states, cities, humans, organizations, academic institutions etc.) were identified in the crawled articles. A *pageRankSqThreshold* of 0.80 was set for pruning the annotations on the basis of their pagerank score.

Evidence Constructor: We implemented Algorithm 1 as a Python script that executes a parametrizable Cypher query to construct candidate evidence sequences; the identified entities in the claim are used as parameters and the concatenated article sections that link these entities together are returned. For our experiments, the maximum number of relationships between the alternating Sections and Entities was set to $2(n-1)$ (shortest path), while the script returns candidate evidence sequences in descending order based on path length. These parameters can be modified if longer candidate evidence sequences are required.

Semantic Similarity: We finetuned a bilingual (Greek-English) *XLM-RoBERTa-base* model (~270M parameters with 12-layers, 768-hidden-state, 3072 feed-forward hidden-state, 8-heads) using 340MB of parallel (EN-EL) sentences from various sources (e.g. OPUS, Wikimatrix, Tatoeba) leveraging the *sentence-transformers* library (Reimers and Gurevych, 2020b). The model was trained for 4 epochs with a batch size of 16 on a machine with a single NVIDIA GeForce RTX3080 (10GB of VRAM) for a total of 28 GPU-hours (single run).

Natural Language Inference: We finetuned a Cross-Encoder *XLM-RoBERTa-base* model of the same architecture on the created Greek-English AllNLI dataset (100MB) using *sentence-transformers*. The model was trained on the same hardware setting for a single epoch, using a train batch size of 6 for 22 GPU-hours (single run).

4.2 Main results

In this section we perform a quantitative and qualitative demonstration of *FarFetched*'s overall performance and also provide individual results for our STS and NLI models based on benchmark datasets.

4.2.1 End-to-end performance

Given the particularity of *FarFetched* in evidence collection (data originating from constantly updating web content), a quantitative evaluation of its performance is quite challenging. To combat the lack of relevant benchmarks for the Greek language, we leveraged the FEVER dataset by Thorne et al. 2018, which models the assessment of truthfulness of written claims as a joint information retrieval and natural language inference task using evidence from Wikipedia. Each row of the dataset comprises a claim in free text, a list of evidence information including a URL to the Wikipedia page of the corresponding evidence and an annotated label (`SUPPORTS`, `REFUTES`, `NOT ENOUGH INFO`). We manually translated a subset of 150 claims from the FEVER validation set from English to Greek and populated the graph database with the content of the corresponding Wikipedia URLs, which was automatically translated into Greek (due to its size), using the NMT model by Papadopoulos et al., 2021. We report *FarFetched*'s performance in terms of accuracy, precision, recall and F1-score on Table 1.

The results indicate a balanced precision and recall for the `REFUTES` and `SUPPORTS` classes,

Label	Precision	Recall	F1-score
`NOT ENOUGH INFO`	.36	.80	.49
`REFUTES`	.91	.72	.80
`SUPPORTS`	.84	.70	.76
Weighted Average	**.82**	**.73**	**.75**
Label accuracy (overall)		**.73**	

Table 1: *FarFetched* claim validation performance on Greek FEVER subset.

while precision is relatively lower for the `NOT ENOUGH INFO` case. This can be partially attributed to the challenges of applying wikification on the automatically translated evidence content, leading to some claims not being linked to their corresponding evidence. Although the above results are not directly comparable to those of similar systems tested on the original English FEVER dataset, they show a significant gain over the baseline model of Thorne et al. 2018 (label accuracy of 0.49). Based on a large comparative study conducted by Bekoulis et al. 2021, *FarFetched* scores in the upper 30th percentile in terms of accuracy (scores ranging from 0.45 to 0.84); however, to the best of our knowledge none of these systems covers the Greek language.

We also provide a set of qualitative examples based on real data that aim at showcasing the capabilities of our system while also acknowledging the dynamicity of the evidence collection process. These scenarios are translated into English to facilitate readability. They include two parts each and are shown in Tables 2, 3 and 4. The original examples (in Greek) are available in the Appendix.

In *Scenario 1*, two contradicting user claims (1a, 1b) with the same entity mentions are provided by the user (Table 2). Since they refer to the same entities, the Evidence Constructor returns the same candidate evidence sequences for both claims in order to evaluate their validity. The most relevant one (STS score in bold) is selected for the NLI phase, where the verdict is that the evidence entails the first claim (1a) and contradicts the second (1b).

In *Scenario 2*, we investigate the sensitivity of our approach in exploiting new information to evaluate a claim (Table 3). The claim initially triggers the Evidence Constructor which returns multiple candidate evidence sequences, in descending STS order (yellow rows). During the NLI evaluation phase, the verdict is entailment, but with a low probability of 0.571 (2a). The same hypothesis is evaluated in Scenario 2b, after the addition of new information appended to the evidence list (blue

User Claim (Scenario 1)	**NLI score**
Denmark and Austria believe that the European Union should increase aid to refugees. (**1a**)	c: 0.014 **e: 0.958** n: 0.028
Denmark disagrees with Austria on the management of immigration issues in the European Union. (**1b**)	**c: 0.951** e: 0.002 n: 0.047
Candidate Evidence Sequences (↓ similarity)	
Austria and Denmark also want to increase EU support for countries hosting refugees near crisis hotspots so that they do not travel to Europe. • **STS Score: 0.8505**	
Checked by police at the Airport Police Departments ... the foreigners presented forged travel documents ... in order to leave the country for other EU countries like France, Germany, Italy, Austria, the Netherlands, Denmark, Spain and Norway. • STS Score: 0.2283	

Table 2: Demonstration of FarFetched on *Scenario 1*.

row). The new evidence is clearly more relevant to the claim at hand, which is successfully identified by *FarFetched*'s STS component that selects it as the best candidate, providing a more confident entailment score of 0.891 (2b). This shift in NLI verdict is visualized in Figure 5. Since *FarFetched* relies on the constantly updating evidence, monitoring such shifts could be useful for identifying trend changes, especially for cases that benefit from long-term planning (business, market, politics etc.)

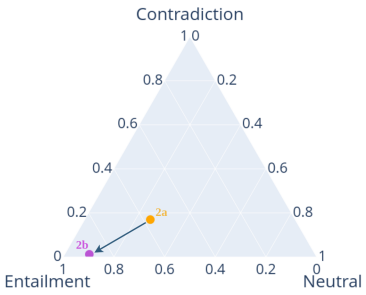

Figure 5: Shift in NLI verdict from Scenario 2a to Scenario 2b of Table 3.

Scenario 3 is similar to *2*, as one claim is evaluated on an initial set of candidate evidence sequences (3a) followed by a new relevant article section with contradicting evidence collected by the Evidence Collector in 3b (Table 4). However, in this case the new evidence is an excerpt from a person's interview. While our approach correctly identifies the relevance of this new evidence to the claim thus affecting the NLI verdict, it does not distinguish between opinions and factual evidence. This is discussed in more detail in Section 5.

User Claim (Scenario 2)	**NLI score** initial (2a)	**NLI score** updated (2b)
The United States plans to impose sanctions on Iran.	c: 0.170 **e: 0.571** n: 0.259	c: 0.012 **e: 0.891** n: 0.097
Candidate Evidence Sequences (↓ similarity)		
Iran faces dilemma over whether to comply of Washington or will lead to collapse. The sanctions that came back in force today, will force the government of the Islamic Republic to accept the US claims regarding the Iranian nuclear program and Iranian activities in the Middle East East because, otherwise, the regime will be in danger to collapse, claimed Israel Kats, the Israeli minister responsible for Information Services. • **STS Score: 0.6665**		
Why Greece was exempted from US sanctions on Iran. New US sanctions on oil exports from Iran have been in force since November 5. • STS Score: 0.6324		
"We are always in favor of diplomacy and talks ... But the Conversations need honesty ... The US is pushing again sanctions on Iran and withdraw from the nuclear deal "(of 2015) and then they want to have conversations with us", Rohani said in a speech that was broadcast live on television. • STS Score: 0.5151		
NEW: Following the collapse of the last talks between the US and Iran, the announcement of additional sanctions is expected in the coming days. • **STS Score: 0.7195**		

Table 3: Demonstration of FarFetched on *Scenario 2*.

User Claim (Scenario 3)	**NLI score** initial (3a)	**NLI score** updated (3b)
Apple is trying to compete with Netflix in the production of television content.	c: 0.004 **e: 0.967** n: 0.029	**c: 0.982** e: 0.008 n: 0.010
Candidate Evidence Sequences (↓ similarity)		
Apple is expected to spend about $ 2 billion this year creating original content that it hopes will compete with Netflix, Hulu and Amazon, already established in the television audience. • **STS Score: 0.7107**		
NEW: "We're not trying to compete with Netflix on TV," an Apple spokesman said in an interview. • **STS Score: 0.7134**		

Table 4: Demonstration of FarFetched on *Scenario 3*.

4.2.2 STS performance

The performance of our semantic similarity model was evaluated on the test subset of the STS2017 benchmark dataset (Cer et al., 2017). Given that the original dataset does not provide sentence pairs in Greek, we manually created a cross-lingual version for the English-Greek pair. The performance is measured using Pearson (r) and Spearman (ρ) correlation between the predicted and gold similarity scores (Table 5). We also provide results regarding translation matching accuracy, evaluating the source and target language embeddings in terms of cosine similarity. Our model achieves a slightly

better performance in both evaluations compared to the current state-of-the-art multilingual model by Reimers and Gurevych, 2019.

Model	STS2017		Translation Matching	
	r	ρ	Acc. (en2el)	Acc. (el2en)
STS-XLM-RoBERTa-base (Ours)	**83.30**	**84.32**	**98.05**	**97.80**
Paraphrase-multilingual-mpnet-base-v2 (UKP-TUDA)	82.71	82.70	97.50	97.35

Table 5: STS model comparison on EN-EL version of STS2017 and in terms of translation matching accuracy.

4.2.3 NLI performance

We benchmark our trained NLI model on the Greek subset of the XNLI dataset (Conneau et al., 2018) that contains 5,010 premise-hypothesis pairs (Table 6). Despite not having used the XNLI dataset during the training phase, we achieve a 1% gain over the multilingual XLM-R (Conneau et al., 2020) and are on par with the monolingual Greek-BERT by Koutsikakis et al., 2020. Since our model was trained on a mixture of Greek and English sentence pairs, it is more suitable for corpora that also contain English terms (e.g. technology, science topics) without suffering from the under-representability of the Greek language occurring in multilingual models.

Model	F1-score
NLI-XLM-RoBERTa-base (Ours)	**78.3**
Greek-BERT (AUEB)	78.6 ± 0.62
XLM-RoBERTa-base (Facebook)	77.3 ± 0.41
M-BERT (Google AI Language)	73.5 ± 0.49

Table 6: NLI model comparison in terms of F1-score on the Greek subset of XNLI-test dataset.

5 Limitations

We acknowledge that *FarFetched* is possible to encounter errors in 3 main areas; these limitations are briefly addressed below.

Entity Linking: Highly ambiguous entities and name variations (e.g. "Washington" could refer to the US state or to "George Washington") pose challenges to any entity linking method. Since we claim that our approach is entity-centric, a wrong entity annotation may lead to irrelevant candidate evidence sequences and increase the probability of "neutral" NLI verdicts. Moreover, the tunable sensitivity of the integrated wikification module implies a trade-off between a precision-oriented and a recall-oriented strategy, the latter resulting in more annotated articles, but also being prone to false-positive annotations.

Evidence Construction: This initial version of our approach relies solely on the STS comparison between the evidence and the claim, based on a shortest path approach as discussed in Section 3.2.4. In cases that involve a larger number of entities in the user claim, calculating the shortest path between the alternating Entity-Section nodes can be computationally cumbersome. Moreover, there is no guarantee that the shortest path is able to capture the most relevant candidate evidence sequences; to this end, outputting the top n best candidates is considered, providing a user with an overview of the extracted news excerpts together with their NLI outcome. Finally, neither a temporal evaluation of the evidence with regard to the claim nor a distinction between opinions and facts is considered; all candidates are treated as equal.

Natural Language Inference: Recognizing the entailment between a pair of sentences partially depends on the tense and aspect of the predications. Tense plays an important role in determining the temporal location of the predication (i.e. past, present or future), while the aspectual auxiliaries signify an event's internal constituency (e.g. whether an action is completed or in progress). While the work of Kober et al., 2019 indicates that language models substantially encode morphosyntactic information regarding tense and aspect, they are unable to reason based only on these properties. To this end, claims with a high presence of such semantic properties should be avoided.

6 Conclusions

In this work, we presented a novel approach for claim validation and reasoning based on the accumulated knowledge from the continuous ingestion and processing of news articles. *FarFetched* is able to evaluate the validity of any arbitrary textual claim by automatically retrieving and aggregating evidence from multiple sources, relying on the pillars of entity linking, semantic textual similarity and natural language inference.

We showcased the effectiveness of our method on the FEVER benchmark as well as on diverse scenarios and acknowledged its limitations. As byproducts of our work, we trained and open-sourced an NLI and an STS model for the less-

resourced Greek language, achieving state-of-the-art performance on the XNLI and STS2017 benchmarks respectively. While our framework fills the gap in automated claim validation for Greek, its modular architecture allows it to be repurposed for any language for which the corresponding models exist.

For future work, we intend to address the limitations of our method mentioned in Section 5, focusing primarily on an optimal entity linking setting, as well as on a more robust strategy for constructing relevant candidate evidence sequences.

Acknowledgments

The research work of Dimitris Papadopoulos was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 50, 2nd call).

References

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017a. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017b. Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, pages 218–223.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Antonin Delpeuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.

William Ferreira and Andreas Vlachos. 2016. 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*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. 2013. Fact checking and analyzing the web. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, page 997–1000, New York, NY, USA. Association for Computing Machinery.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. ForecastQA: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4636–4650, Online. Association for Computational Linguistics.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sarthak Majithia, Fatma Arslan, Sumeet Lubal, Damian Jimenez, Priyank Arora, Josue Caraballo, and Chengkai Li. 2019. ClaimPortal: Integrated monitoring, searching, checking, and analytics of factual claims on Twitter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 153–158, Florence, Italy. Association for Computational Linguistics.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

Dimitris Papadopoulos, Nikolaos Papadakis, and Nikolaos Matsatsinis. 2021. PENELOPIE: Enabling open information extraction for the Greek language through machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 23–29, Online. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning to predict from textual data. *Journal of Artificial Intelligence Research*, 45:641–684.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020b. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mehdi Samadi, Partha Talukdar, Manuela Veloso, and Manuel Blum. 2016. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2019. Learning to update knowledge graphs by reading news. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2632–2641, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction.

In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, et al. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.

Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. 2015. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49, Beijing, China. Association for Computational Linguistics.

Jim Webber. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH '12, page 217–218, New York, NY, USA. Association for Computing Machinery.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qi Zeng, Manling Li, Tuan Lai, Heng Ji, Mohit Bansal, and Hanghang Tong. 2021. GENE: Global event network embedding. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 42–53, Mexico City, Mexico. Association for Computational Linguistics.

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. *FaxPlainAC: A Fact-Checking Tool Based on EXPLAINable Models with HumAn Correction in the Loop*, page 4823–4827. Association for Computing Machinery, New York, NY, USA.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

A Appendix: Original examples (in Greek) of Tables 2, 3 and 4.

User Claim (Scenario 1)	**NLI score**
Η Δανία και η Αυστρία πιστεύουν ότι η Ευρωπαϊκή Ένωση πρέπει να αυξήσει τη βοήθεια προς τους πρόσφυγες. (**1a**)	c: 0.014 **e: 0.958** n: 0.028
Η Δανία διαφωνεί με την Αυστρία σχετικά με τη διαχείριση των μεταναστευτικών θεμάτων στην Ευρωπαϊκή Ένωση. (**1b**)	**c: 0.951** e: 0.002 n: 0.047
Candidate Evidence Sequences (↓ similarity)	
Η Αυστρία και η Δανία θέλουν να ενισχυθεί επίσης η υποστήριξη της ΕΕ προς κράτη που υποδέχονται πρόσφυγες κοντά σε εστίες κρίσεις, ώστε οι πρόσφυγες αυτοί να μην ταξιδεύουν προς την Ευρώπη. • **STS Score: 0.8505**	
Σε έλεγχο από αστυνομικούς των Αστυνομικών Τμημάτων Αερολιμένων ... οι αλλοδαποί επέδειξαν πλαστά ταξιδιωτικά έγγραφα προκειμένου να αναχωρήσουν από τη χώρα για άλλες χώρες της ΕΕ όπως η Γαλλία, Γερμανία, Ιταλία, Αυστρία, Ολλανδία, Δανία, Ισπανία και Νορβηγία. • STS Score: 0.2283	

Table 7: Demonstration of FarFetched on *Scenario 1* in Greek language.

User Claim (Scenario 2)	**NLI score** initial (2a)	**NLI score** updated (2b)
Οι Ηνωμένες Πολιτείες σχεδιάζουν να επιβάλλουν κυρώσεις στο Ιράν.	c: 0.170 **e: 0.571** n: 0.259	c: 0.012 **e: 0.891** n: 0.097
Candidate Evidence Sequences (↓ similarity)		
Το Ιράν μπροστά στο δίλημμα αν θα συμμορφωθεί προς τις υποδείξεις της Ουάσινγκτον ή θα οδηγηθεί σε κατάρρευση Οι κυρώσεις που επανήλθαν σε ισχύ σήμερα, θα αναγκάσουν την κυβέρνηση της Ισλαμικής Δημοκρατίας να δεχθεί τις αξιώσεις των ΗΠΑ όσον αφορά το ιρανικό πυρηνικό πρόγραμμα και τις ιρανικές δραστηριότητες στην περιοχή της Μέσης Ανατολής διότι, σε διαφορετική περίπτωση, το καθεστώς θα κινδυνεύσει να καταρρεύσει, υποστήριξε ο Ισραέλ Κατς, ο ισραηλινός υπουργός αρμόδιος για τις Υπηρεσίες Πληροφοριών. • **STS Score: 0.6665**		
Γιατί εξαιρέθηκε η Ελλάδα από τις αμερικανικές κυρώσεις στο Ιράν. Από τις 5 Νοεμβρίου βρίσκονται σε ισχύ οι νέες κυρώσεις των ΗΠΑ για εξαγωγές πετρελαίου από το Ιράν. • STS Score: 0.6324		
«Είμαστε πάντα υπέρ της διπλωματίας και των συνομιλιών ... Όμως οι συνομιλίες χρειάζονται εντιμότητα ... Οι ΗΠΑ επιβάλλουν εκ νέου κυρώσεις στο Ιράν και αποσύρονται από την πυρηνική συμφωνία (του 2015) και μετά θέλουν να κάνουν συνομιλίες μαζί μας», δήλωσε ο Ροχανί σε ομιλία του που μεταδόθηκε ζωντανά από την τηλεόραση. • STS Score: 0.5151		
Μετά το ναυάγιο των τελευταίων συνομιλιών μεταξύ ΗΠΑ και Ιράν αναμένεται η ανακοίνωση επιπλέον κυρώσεων τις επόμενες ημέρες. • **STS Score: 0.7195**		

Table 8: Demonstration of FarFetched on *Scenario 2* in Greek language.

User Claim (Scenario 3)	**NLI score** initial (3a)	**NLI score** updated (3b)
Η Apple προσπαθεί να ανταγωνιστεί την Netflix στην παραγωγή τηλεοπτικού περιεχομένου.	c: 0.004 **e: 0.967** n: 0.029	**c: 0.982** e: 0.008 n: 0.010
Candidate Evidence Sequences (↓ similarity)		
Η Apple αναμένεται να δαπανήσει φέτος περίπου 2 δισεκατομμύρια δολάρια με σκοπό τη δημιουργία πρωτότυπου περιεχομένου που ελπίζει ότι θα ανταγωνιστεί τις ήδη εδραιωμένες στο τηλεοπτικό κοινό υπηρεσίες των Netflix, Hulu και Amazon. • **STS Score: 0.7107**		
«Δεν προσπαθούμε να ανταγωνιστούμε το Netflix στην τηλεόραση», δήλωσε εκπρόσωπος της Apple σε συνέντευξή του. • **STS Score: 0.7134**		

Table 9: Demonstration of FarFetched on *Scenario 3* in Greek language.