

*Web Science – Investigating the Future of Information and Communication*

# Stance Classification for Fact Checking

**Pavlos Fafalios**

fafalios@L3S.de

May 2019

**Web Science 2019**

# Fake news

*“A type of yellow journalism or propaganda that consists of deliberate disinformation or hoaxes spread via traditional print and broadcast news media or online social media.” – Wikipedia*

- Fabricated news that takes the appearance of real news
- Increasingly prevalent over the last few years
- Difficult to detect



# Fake News



## Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement – WTOE 5 News

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement TOPICS:Pope Francis endorses Donald Trump photo by Jeffrey Bruno /...

## Trump IQ test results discovered in former NYMA employee's closet

*The result: 73.*

The results of an IQ test that President Donald Trump allegedly took during his first year at New York Military Academy have been discovered in a file box in a closet in Brooklyn. According to the test results, Trump's IQ is 73.

The document, currently in the process of being authenticated, was discovered Thursday by William Askew, Jr. as he was cleaning out his late father's apartment.

"Dad was the school counselor at NYMA from 1955 to 1985. He didn't administer these tests, but he was in charge of collecting them and sending them to the grading office," said Askew.

Askew produced additional documents and photographs that confirm his father was indeed employed as school counselor at NYMA at the same time Trump attended the military prep school.

"I don't think this should be news to anybody," said Askew. "The guy's clearly a total moron."



Askew today, and his late father at work at NYMA.

21 hours ago

## Michelle Obama Deletes Hillary Clinton From Twitter

When Hillary goes low, Michelle goes BYE!

Posted on November 1, 2016 by Baxter Dmitry in News, US // 19 Comments



## Black Lives Matter THUGS Blocking Emergency Crews From Reaching Hurricane Victims

LOCK THEM ALL UP! THIS IS NO TIME FOR GAMES!

© August 26, 2017 David "Tango" Foxtrot



Pavlos Fafalios, Stance Classification for Fact Checking, Web Science 2019



INCIDENTS

THEREISNEWS.com

## Two altar boys were arrested for putting weed in the censer-burner

# Fake news – Types *(by Claire Wardle)*

- fabricated content
  - new content is 100% false, designed to deceive and do harm
- misleading content
  - misuse of information to frame an issue completely differently
- false context
  - when genuine content is shared with false contextual information
- impostor content
  - Fake content that purports to come from a real news site
- manipulated content
  - when genuine information or imagery is manipulated to deceive
- false connection
  - when headlines, visuals or captions don't support the content
- satire or parody
  - no intention to cause harm but has potential to fool

# Fake news – Influence of Social Media

- Twitter:
  - False claims are retweeted faster, further, and for longer than true claims (Vosoughi et al. 2018)
- Facebook:
  - The top-20 fake news stories about 2016 US Election received more engagement than the top-20 election stories from 19 major media outlets  
<https://abcnews.go.com/Technology/fake-news-stories-make-real-news-headlines/story>
  - 50 of the biggest fake stories of 2018 generated roughly 22 million total shares, reactions, and comments on Facebook  
<https://www.buzzfeednews.com/article/craigsilverman/facebook-fake-news-hits-2018>

Vosoughi, S., Deb R., and Sinan A. "The spread of true and false news online." *Science* 359, no. 6380 (2018): 1146-1151.

# Fact checking

- An assertion or statement (a “**claim**”) is examined to determine its veracity and correctness
  - Ante hoc: before publication/dissemination (for publishing only checked material)
  - Post hoc: after publication/dissemination (for checking the veracity of a published claim)
- Fact-checking organizations: from 44 in 2014 to almost 100 in 2016  
<https://reporterslab.org/global-fact-checking-up-50-percent/>
- Popular “**post hoc**” fact-checking web sites:
  - FactCheck.org - 4 Webby Awards in the Politics category
  - PolitiFact - Pulitzer Prize for National Reporting in 2009
  - Snopes - *"well-regarded source for sorting out myths and rumors"*
- High impact (Nyhan and Reifler 2015)
  - corrective effect on misperceptions among citizens
  - discourages politicians from spreading misinformation

Nyhan, B., & Reifler, J. (2015). *The effect of fact-checking on elites: A field experiment on US state legislators*. American Journal of Political Science, 59(3), 628-640.

# Fact checking - Example

<https://www.snopes.com/fact-check/donald-trump-nyma-iq/>

The screenshot shows the Snopes website interface. At the top, there is a yellow navigation bar with the Snopes logo and a search bar. Below the navigation bar, there are links for 'Submit a Topic', 'Shop Snopes', 'What's New', 'Hot 50', 'Fact Checks', 'News', 'Videos', 'Archive', and 'About'. The main content area features a breadcrumb trail 'Fact Check > Politics'. The article title is 'Was Donald Trump's IQ Measured at 73?'. The sub-headline reads 'Trump's IQ test results were supposedly discovered in a former New York Military Academy employee's closet.' The author is 'DAN EVON' and the publication date is 'PUBLISHED 6 MAY 2019'. Below the text is a photograph of Donald Trump sitting in a yellow chair, wearing a dark suit and a red tie.

## Claim

A document discovered in May 2019 reveals that Donald Trump's IQ was measured at 73 during his high school years.

## Rating



**False**

*About this rating* [↗](#)

## Origin

The IQ of U.S. President Donald Trump has been a subject of speculation since the early days of the 2016 presidential election cycle. During that campaign, for example, we investigated a [false rumor](#) that Trump had the second highest IQ (156) of all **presidents** in U.S. history. In May 2019, we encountered a rumor claiming the opposite: that Trump's IQ was a lowly 73 (placing him in the "well below average" classification).

This rumor circulated in the form of an image that showed a purported newspaper clipping about the recent discovery of information regarding an IQ test Trump allegedly took during his first year at New York Military Academy (NYMA):

### Trump IQ test results discovered in former NYMA employee's closet

**The result: 73.**

The results of an IQ test that President Donald Trump allegedly took during his first year at New York Military



# Fact checking – Topics and Ratings



Fact Check Archives

|                 |                     |                   |
|-----------------|---------------------|-------------------|
| Automobiles     | Business            | Cokelore          |
| College         | Computers           | Crime             |
| Critter Country | Disney              | Embarrassments    |
| Entertainment   | Junk News           | Fauxtography      |
| Food            | Fraud & Scams       | Glurge Gallery    |
| History         | Holidays            | Horrors           |
| Humor           | Hurricane Katrina   | Inboxer Rebellion |
| Language        | Legal Affairs       | Lost Legends      |
| Love            | Luck                | Media Matters     |
| Medical         | Military            | Old Wives' Tales  |
| Politics        | Questionable Quotes | Racial Rumors     |
| Religion        | Risqué Business     | Science           |
| September 11th  | Sports              | Superstition      |
| Technology      | Travel              | Viral Phenomena   |
| Weddings        | Soapbox             | War/Anti-War      |

**Verdict (Rating)**

Fact Checks by Rating

|               |              |                     |
|---------------|--------------|---------------------|
| True          | Mostly True  | Mixture             |
| Mostly False  | False        | Unproven            |
| Outdated      | Miscaptioned | Correct Attribution |
| Misattributed | Scam         | Legend              |



# Fact checking

- Challenging process:
  - Laborious, demanding, time-consuming, costly
    - 1 day to research and write a typical article about a single claim (Hassan et al. 2015)
  - Difficult to keep up with the amount of misinformation and the spread speed
  - Lack of resources for investigative journalism!
- Facilitating fact-checking
  - Trying to (semi-)automate some of its stages:
    - Detecting check-worthy claims
    - Matching check-worthy claims to fact-checked claims
    - Finding claim-relevant documents
    - Detecting the stance of a document towards a claim
    - Inferring the veracity of a claim

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., & Yu, C. (2015, July). The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium*.

# Stance Detection

- Detecting the perspective (stance) of a document towards a claim
  - It provides evidence to support true claims or invalidate false claims (for facilitating a fact-checking process)
  - It provides a means to identify potential misinformation (if the claim has already been fact-checked)
- Input:
  - The text of a **claim**
  - The text of a **document**
- Output:
  - The **stance** of the document towards the claim
    - UNRELATED
    - DISCUSS (NEUTRAL)
    - AGREE
    - DISAGREE

A multi-class  
classification  
problem!

# Stance Detection – Example

- Claim: ***Graffiti Artist Banksy Arrested In London; Identity Revealed***

 **Jack Monroe**  
@MsJackMonroe

Just heard the news about [#Banksy](#) - a sad day for politics, freedom, creativity and protest. Disappointed with media for publishing his ID.

← Reply ↻ Retweet ★ Favorite ⋮ More

 **Ross Worswick** ✓  
@rossworswick

Soo the police have arrested the genius that is Banksy... Well done guys the streets are a safer place now 🙌😂

♡ 47 1:19 PM - Oct 20, 2014

💬 23 people are talking about this

 **Jo Brooks**  
@brightonseagull

I have some bad news everyone. Israeli Police arrested Banksy early this morning, still trying to get more info. Will post when I know more.

2:51 AM - 17 Sep 2017

6 Retweets 8 Likes

💬 6 ♡ 8

 **follow @thunderclap**  
@billyrayanus

banksy's art isnt mindless vandalism, its art put there to provoke thought in people and he doesnt deserve to be arrested for making art

♡ 12 12:55 PM - Oct 20, 2014

👤 See follow @thunderclap's other Tweets

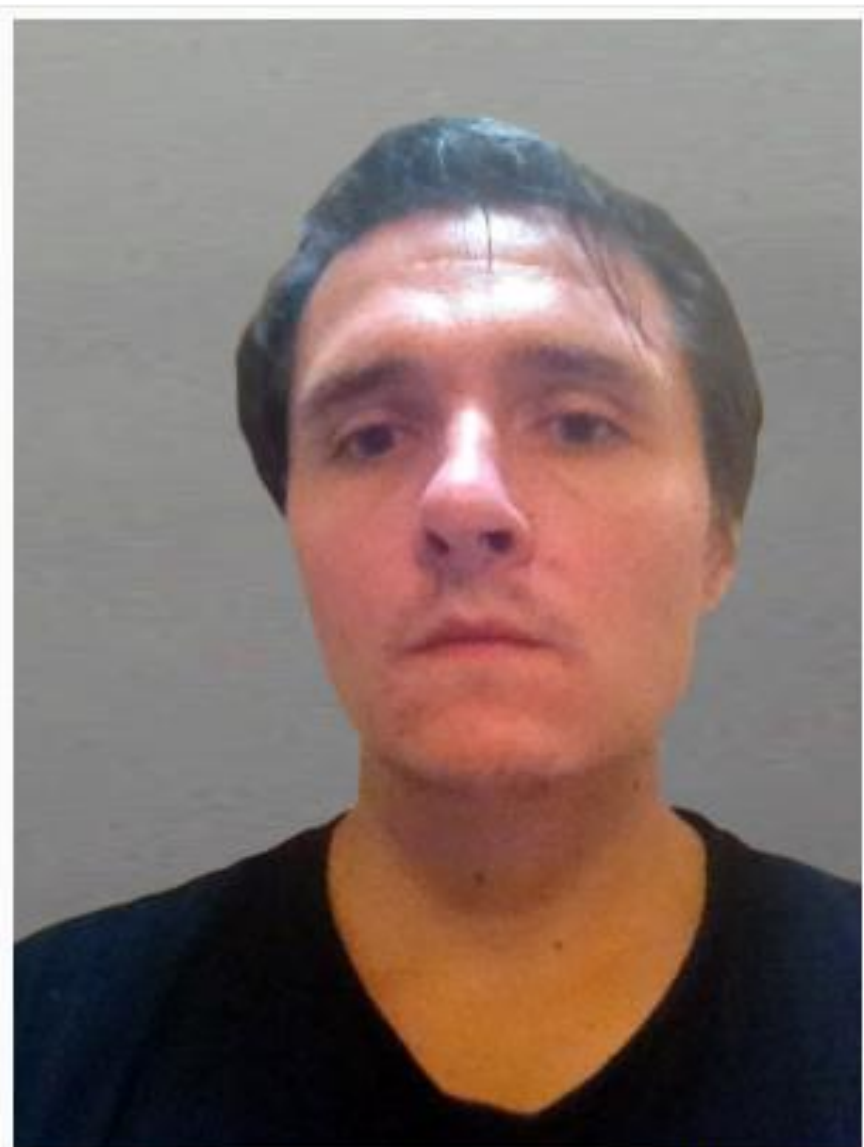
# Stance Detection – Example

- Document:

Stance: **AGREE**

## Graffiti Artist Banksy Arrested In London; Identity Revealed

Posted about 6 days ago | 396 comments



A mugshot taken by London Police today of Paul Horner AKA Banksy. (AP Photo/Dennis System, File) / AP

London, England — The elusive graffiti artist, political activist, film director, painter and long time fugitive that for years has gone by the pseudonymous name of Banksy, was arrested early this morning by London's Metropolitan Police. After hours of questioning and a raid of his London art studio, his true name and identity have finally been revealed.

The City of London Police say Banksy's real name is Paul Horner, a 35-year old male born in Liverpool, England. The BBC has confirmed this information with Banky's PR agent Jo Brooks along with Pest Control, a website that acts as a handling service on behalf of the artist.

London Police Chief Lyndon Edwards held a press conference to answer questions about Banksy and how Horner was finally apprehended after all these years on the run.

"We had a 24-hour Anti-Graffiti Task Force monitoring different groups where Banksy was known to frequent. We received word that around 2am a group of individuals left a flat speculated to be one of Banky's art studios. This group was followed by agents and once vandalism had occurred, we then arrested the group, 5 men total. These individuals all had ID on them except for one, and that is the one we believed to be Banksy," Edwards said. "We then raided the studio where the group was last seen leaving from. Inside we found thousands of dollars of counterfeit money along with future projects of vandalism. We also found


# Stance Detection – Example

- Document:

Stance: **DISAGREE**

**Banksy 'Arrested & Real Identity Revealed'**

Sara C Nelson  
Huffington Post UK



Elusive graffiti artist Banksy's cover was blown when he was unceremoniously arrested for vandalism, conspiracy, racketeering and counterfeiting.

Well that's what US website the National Report would have you believe.

The bogus story alleged the infamous street artist was arrested following a raid on his London studio.

It 'outs' him as Paul Horner, a 34-year-old born in Liverpool and says he is being

**SUBSCRIBE AND FOLLOW**  
Get top stories and blog posts emailed to you each day. Newsletters may offer personalized recommendations and advertisements. [Learn more](#)

✉ **Newsletter**

*address@email.com*

[Subscribe Now](#) →

# Stance Detection – Example

- Document:

TRENDING Published January 8, 2015 By Staff Reporter

## **Banksy Reportedly Arrested!**

Reports of the famous graffiti artist Banksy having been arrested in London has been making the rounds on the internet as of late.

A website called "National Report" have published an article saying that the modern-day artist has finally been arrested and his identity has even been revealed. They wrote that Banksy was identified as a man named Paul Horner, thirty-five years of age. Aside from his identity, the article also said that Banksy has been arrested by both the Metropolitan Police, and the City of London Police. This caused numerous backlashes from the public, bashing the London police for

Stance: **DISCUSS**

# Stance Detection – Example

- Document:

Stance: **UNRELATED**

## New home for Welsh Banksy work agreed

🕒 30 April 2019

f 🗨️ 🐦 ✉️ Share



Thousands of people have visited the Banksy piece since it appeared in December.

An agreement on where Port Talbot's Banksy will go on display has been reached ahead of work to move the graffiti art to a new home.

Neath Port Talbot Council said it has agreed to let a space at the town's Ty'r Orsaf buildings to house the 'Season's Greetings' piece.

Pavlos The work **was bought by art dealer John Brandler** for a six-figure sum with the promise to put it on public show.

# Stance Detection – Fake News Challenge (FNC)

- Fake News Challenge (FNC) - <http://www.fakenewschallenge.org/>
  - *“The **goal** of the Fake News Challenge is to explore how artificial intelligence technologies, particularly **machine learning** and **natural language processing**, might be leveraged to **combat the fake news problem**”*
  - *“A helpful first step towards identifying fake news is to **understand what other news organizations are saying about the topic**”*
- **FNC-I task: Stance Detection**
  - Estimating the stance of a body text from a **news article** (“document”) relative to a **headline** (“claim”)
  - The body text may **AGREE, DISAGREE, DISCUSS** or be **UNRELATED** to the headline



# Stance Detection – Fake News Challenge (FNC)

- Dataset

- Derived from the Emergent project (Ferreira and Vlachos 2016)
- 2,587 documents related to 300 claims (200 for training and 100 for testing)
- Each document has a summarized “headline” that reflects its stance
  - Thus, each claim is represented through different headlines of different stances!

- Example:

**Headline (claim):** *“No, it's not Tiger Woods selling an island in Lake Mälaren”*

**Document:** *“The sale of a private island in Lake Mälaren has received international attention, thanks to the assertion that it is Tiger Woods who sells it. But that's not true - the true owner is a Swedish millionaire ... ..”*

**Stance:** AGREE

# Stance Detection – Fake News Challenge (FNC)

- Instances (claim-document pairs):

|       | All    | Unrelated | Neutral | Agree | Disagree |
|-------|--------|-----------|---------|-------|----------|
| Train | 49,972 | 36,545    | 8,909   | 3,678 | 840      |
| Test  | 25,413 | 18,349    | 4,464   | 1,903 | 697      |

**Highly unbalanced  
class distribution!**

- **AGREE** and **DISAGREE**: important classes in this “fake news” context
  - Discovery of documents that can help invalidating false claims (e.g., by providing evidence)
  - Discovery of sources that distribute fake news!

# Stance Detection: Approaches

# Stance Detection: FNC-I baseline

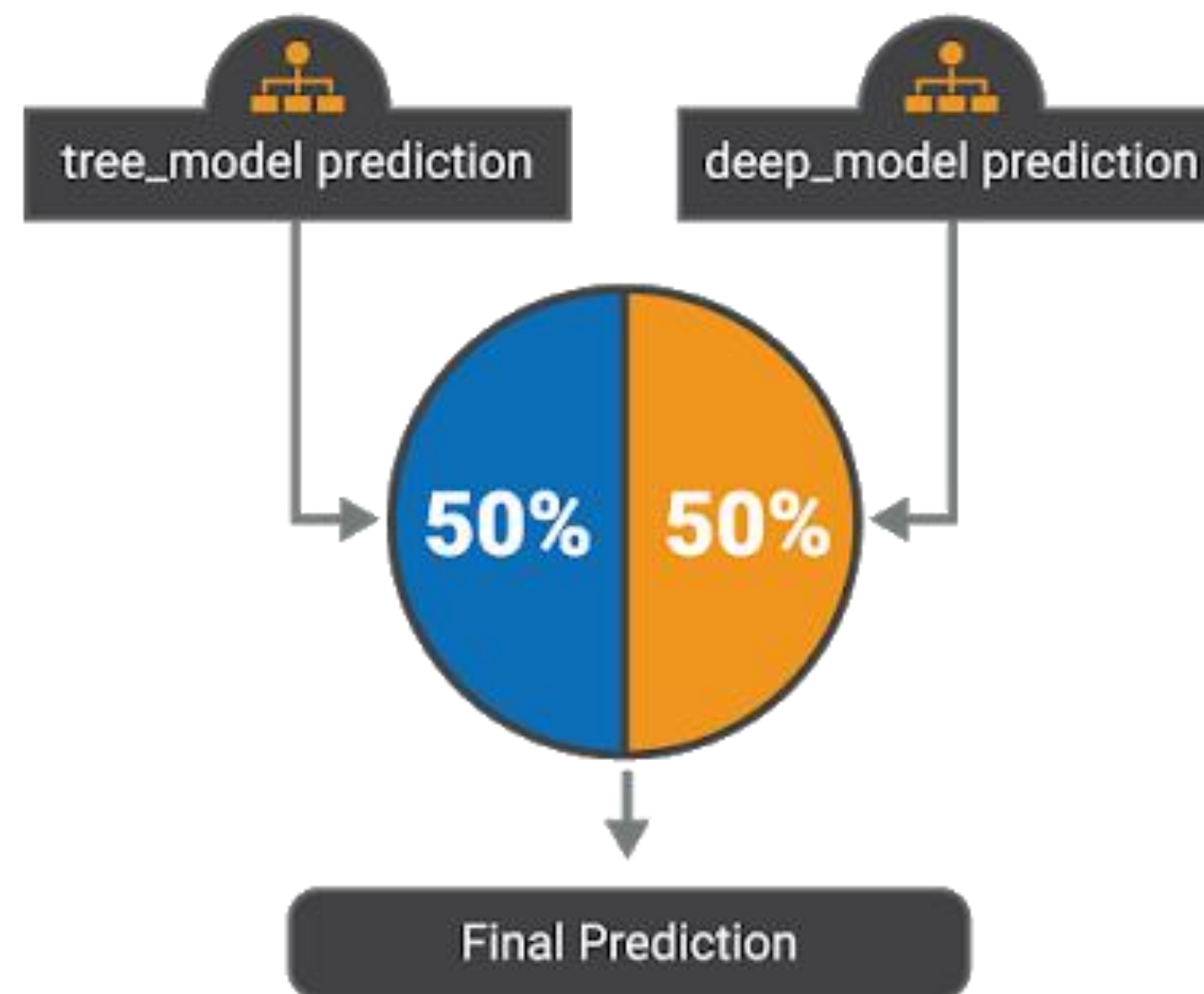
<https://github.com/FakeNewsChallenge/fnc-1-baseline>

- **Gradient boosting classifier**
- **Features:**
  - **N-grams match:** number of common  $n$ -grams (sequence of  $n$  continuous words) in the headline and the document (2-, 3-, 4-, 5-, 6-grams)
  - **N-chargrams match:** Number of common  $n$ -chargrams (sequence of  $n$  continuous characters) in the headline and the document (2-, 8-, 4-, 16-chargrams)
  - **Binary co-occurrence:** number of words of the claim that appear i) in the entire body of the document, ii) in the first 255 words of the document
  - **Lemma overlap:** similar to unigram match, but the words are first lemmatized.
  - **Refuting words:** Appearance of refuting words in the headline
    - List: 'fake', 'fraud', 'hoax', 'false', 'deny', 'denies', 'not', 'despite', 'nope', 'doubt', 'doubts', 'bogus', 'debunk', 'pranks', 'retract'
  - **Headline polarity:** number of refuting words in headline % 2
  - **Document polarity:** number of refuting words in the document % 2

# Stance Detection: Solat in the SWEN

<https://github.com/Cisco-Talos/fnc-1/>

- The top-ranked system of FNC-I
  - Combination (weighted average) between **gradient boosting** and a **deep convolutional neural network**



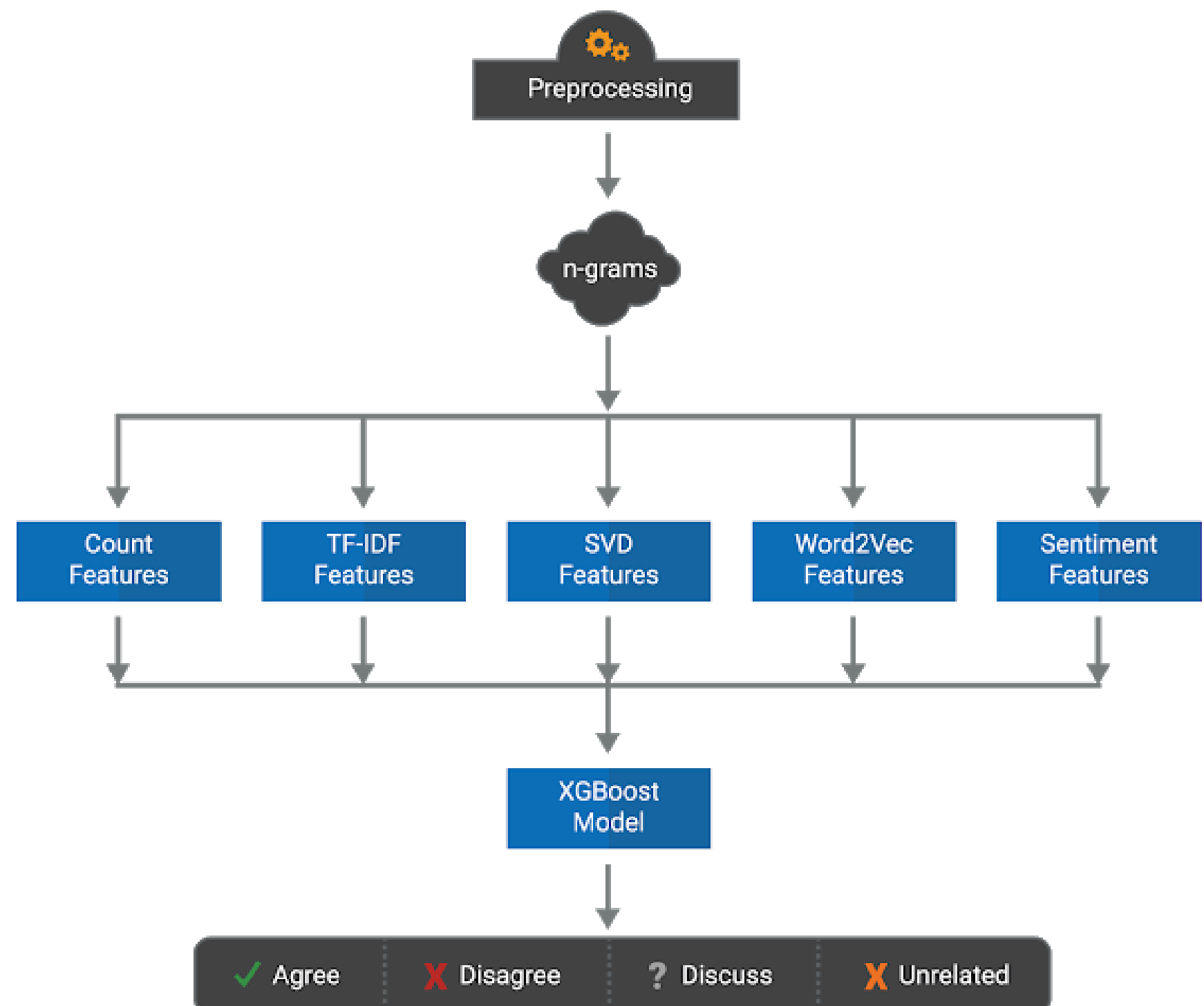
# Stance Detection: Solat in the SWEN

<https://github.com/Cisco-Talos/fnc-1/>

- **Gradient boosted decision trees**

- **Features:**

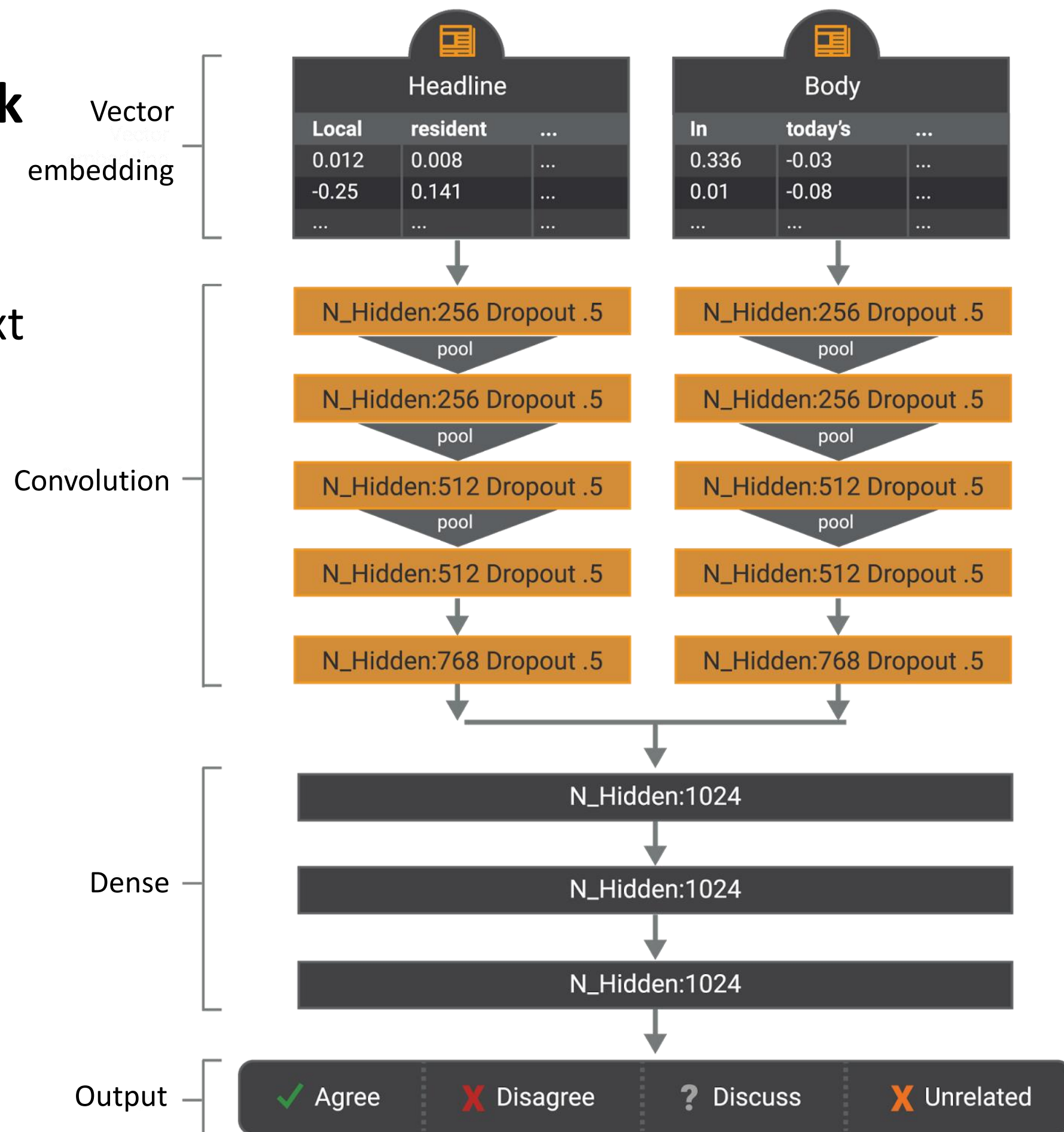
- Basic Count Features
- TF-IDF Features
- SVD Features
  - For finding the latent topics
- Word2Vec Features
  - Using pre-trained word vectors
- Sentiment Features
  - Sentiment polarity score of headline and document
  - Using NLTK (<https://www.nltk.org/>)



# Stance Detection: Solat in the SWEN

<https://github.com/Cisco-Talos/fnc-1/>

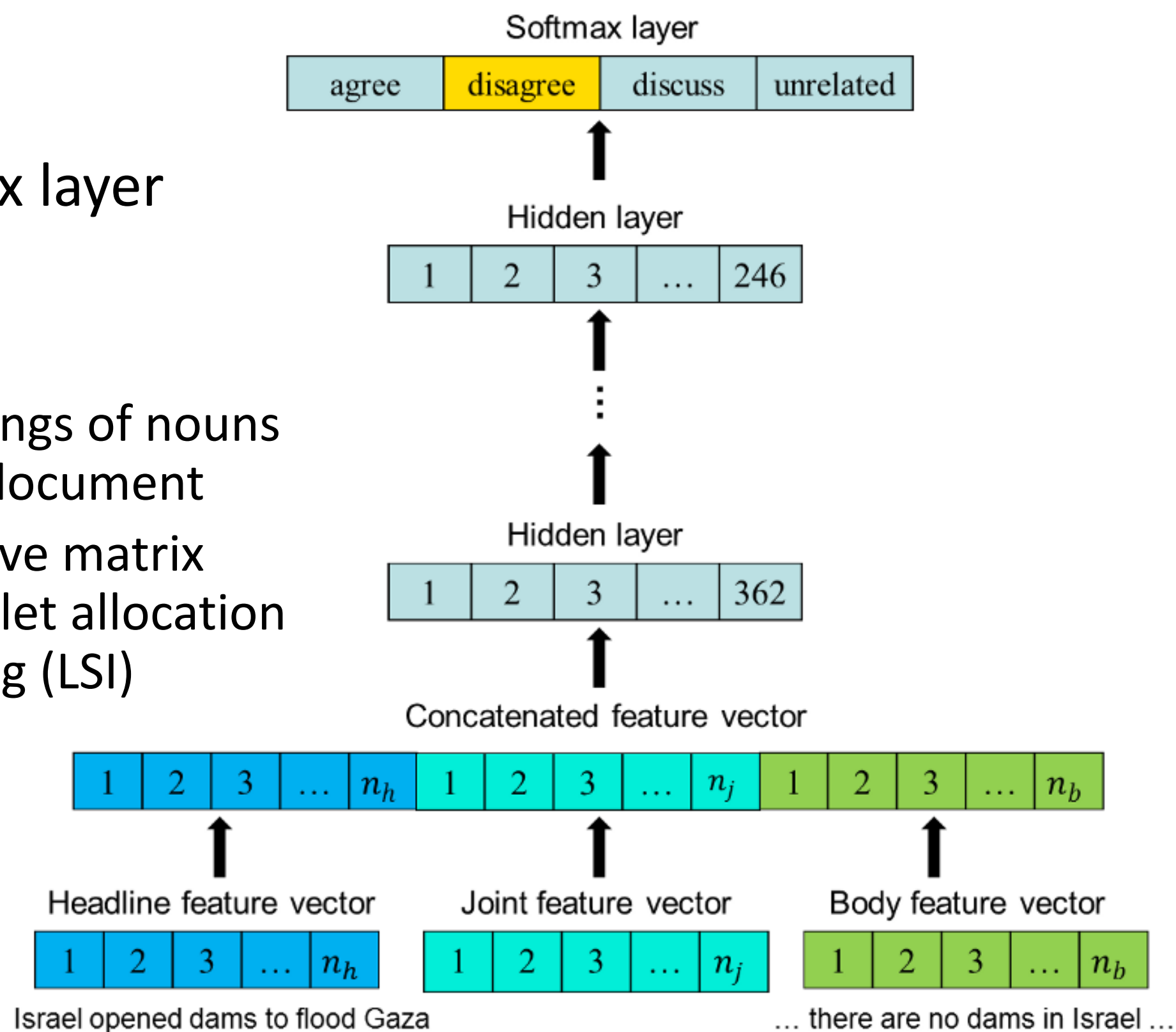
- **Deep convolutional neural network**
- Networks:
  - 1D CNN on the headline and body text using pre-trained word2vec
  - MLP with 4-class output



# Stance Detection: Athene (UKP Lab)

[https://github.com/hanselowski/athene\\_system](https://github.com/hanselowski/athene_system)

- The 2<sup>nd</sup> ranked system of FNC-I
- MLP with 6 hidden and a softmax layer
- Features:
  - Unigrams
  - Cosine similarity of word embeddings of nouns and verbs between headline and document
  - Topic models based on non-negative matrix factorization (NNMF), latent dirichlet allocation (LDA), and latent semantic indexing (LSI)
  - + the FNC-I baseline features

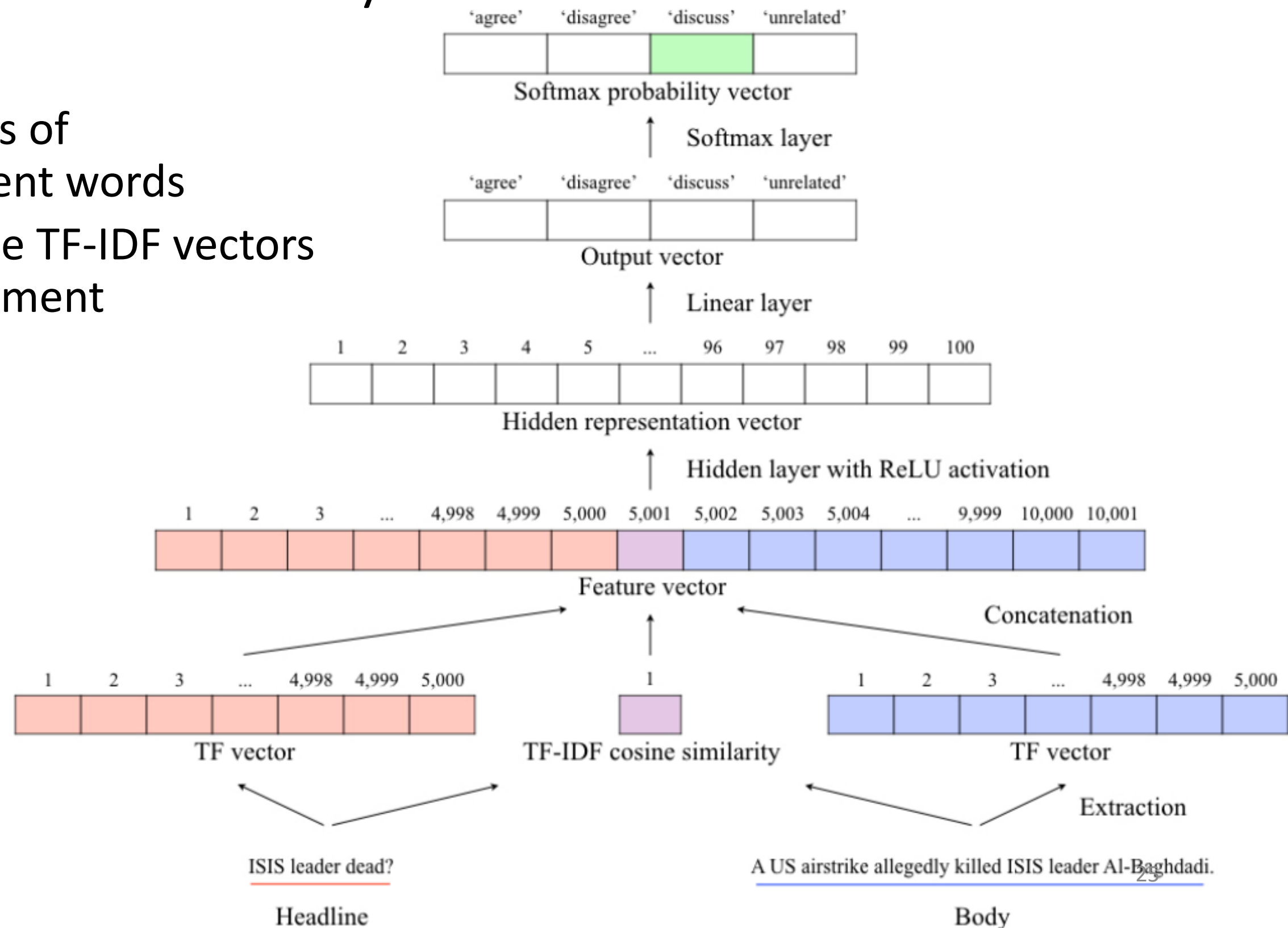




# Stance Detection: UCL Machine Reading

<https://github.com/uclmr/fakenewschallenge>

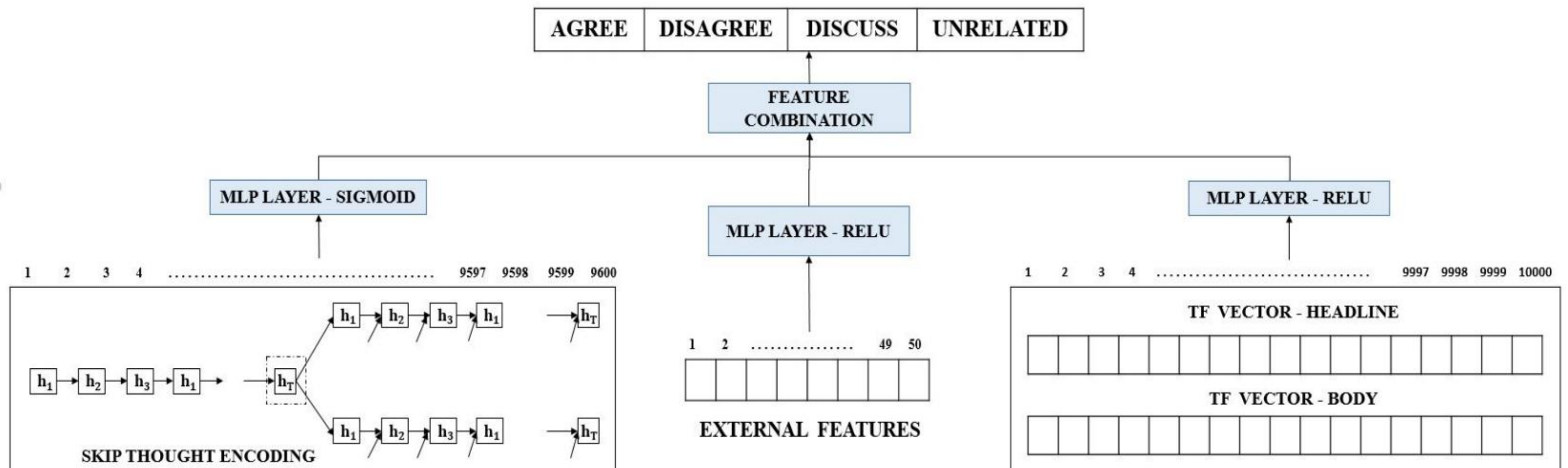
- The 3<sup>rd</sup> ranked system of FNC-I
- Simple MLP network with 1 hidden layer
- Features:
  - TF vectors of unigrams of the 5,000 most frequent words
  - Cosine similarity of the TF-IDF vectors of the claim and document



# Stance Detection: CombNSE

Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., & Mittal, A. (2018). **Combining Neural, Statistical and External Features for Fake News Stance Identification.** In *Companion of the The Web Conference 2018*.

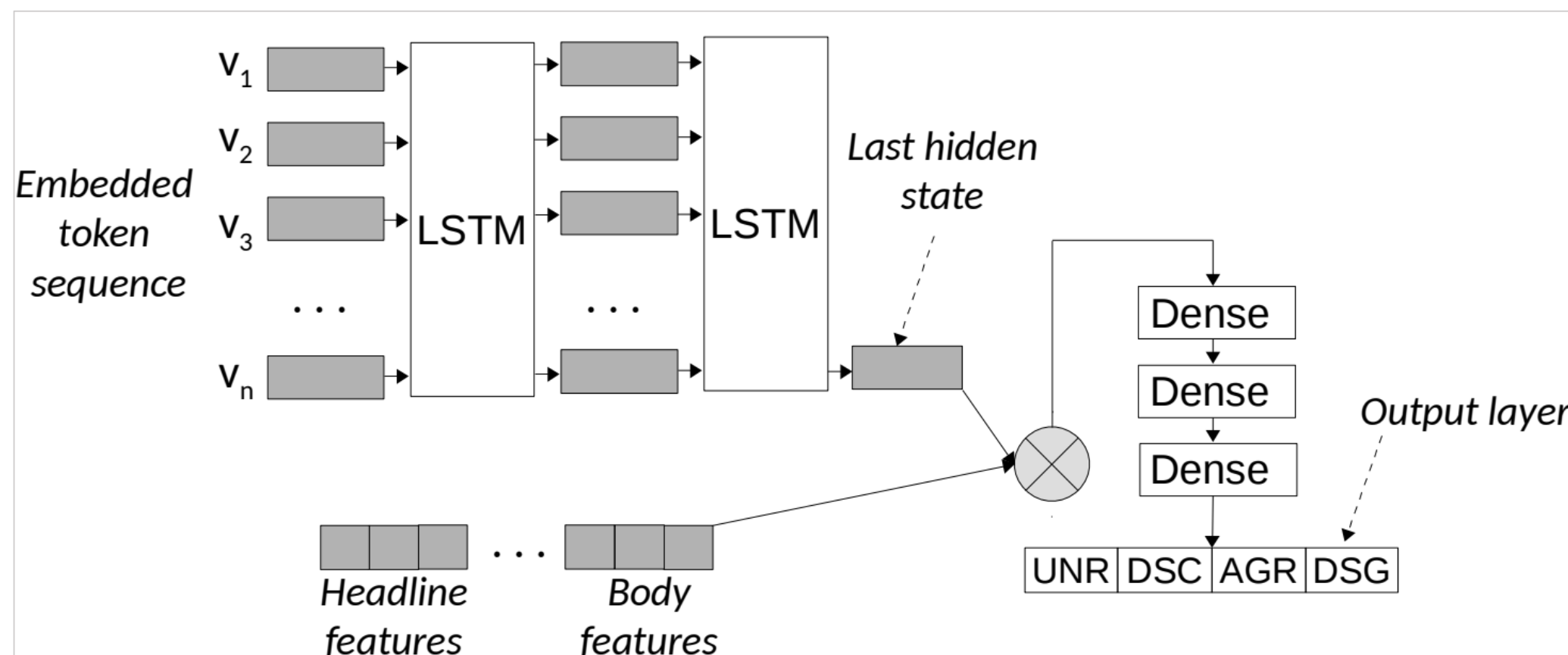
- Deep MLP model combining neural, statistical and external features
- Neural features
  - Skip-thought vectors which encode sentences to vector embedding
- Statistical features
  - 1-gram TF vector of the headline, 1-gram TF vector of the body
- External features
  - $n$ -grams ( $n = 2, \dots, 6$ ),  $n$ -chagrams ( $n = 2, \dots, 16$ ), TF-IDF
  - Sentiment difference between the headline-body pair (using a lexicon based approach)



# Stance Detection: StackLSTM

Hanselowski, A., Avinesh, P. V. S., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). **A Retrospective Analysis of the Fake News Challenge Stance-Detection Task**. In *COLING 2018*.

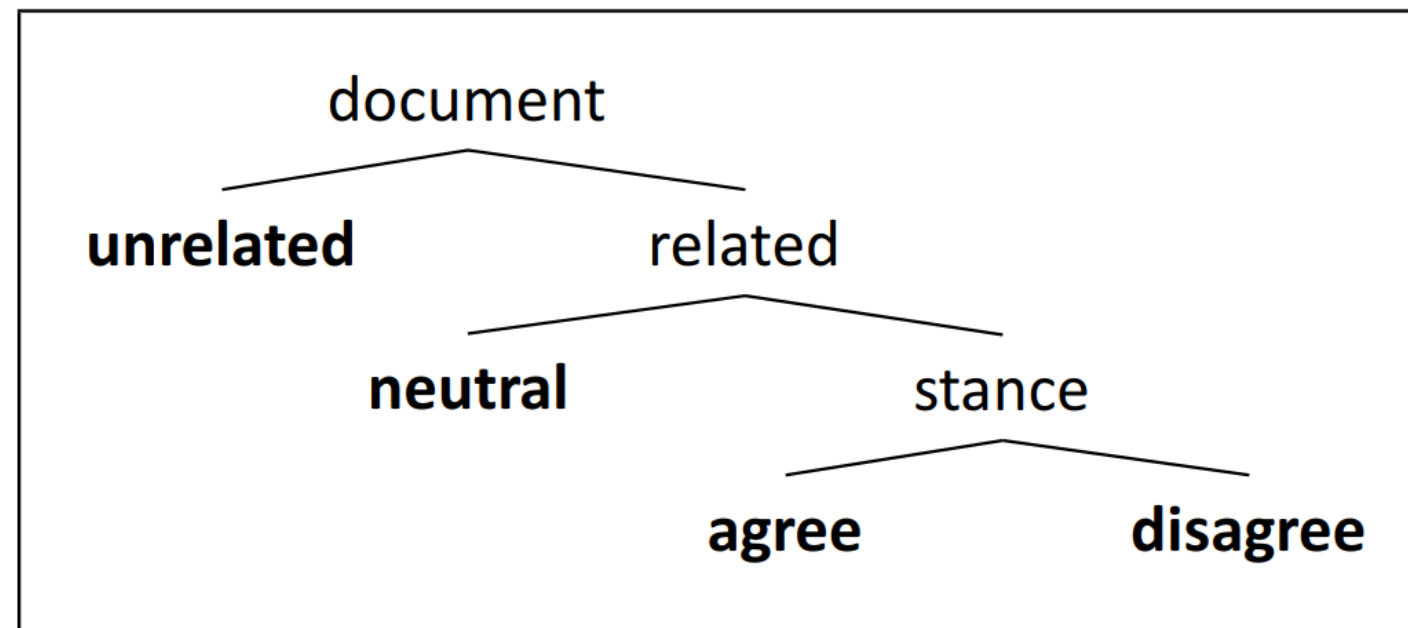
- Stacked LSTM network (RNN) combined with a set of hand-crafted features
  - Using 50-dimensional GloVe word embeddings (Pennington et al. 2014) for generating sequences of word vectors of a headline-document pair
    - *For better capturing the meaning of the sentence*
- Features:
  - BoW unigram features
  - Bag-of-character 3-grams features
  - Topic model based features (the ones used in *Athene* system)



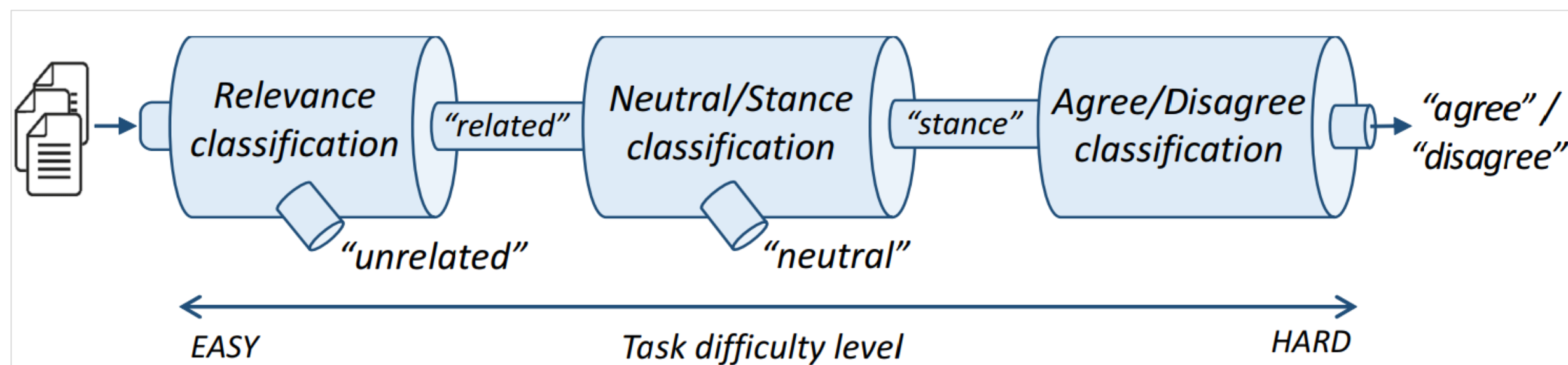
# Stance Detection: L3S (Learning in 3 Steps)

*Under review...*

- Observation: tree-like hierarchy of the 4 classes



- 3-stage pipeline



# Stance Detection: L3S (Learning in 3 Steps)

- **Stage 1: Relevance classification**

- **SVN classifier with class-wise penalty**
- Lexical features:
  - n-grams match, chargrams match, binary co-occurrence, lemma overlap, cosine similarity, word2vec similarity, **keyword** overlap, **proper noun** overlap



- **Stage 2: Neutral/Stance classification**

- Simple **CNN model** with embedded word vectors and sentiment features
- Sentiment features:
  - Using NLTK (<https://www.nltk.org/>) on first 10 sentences (output: array of 4 sentiment scores: positive, negative, neutral, compound)



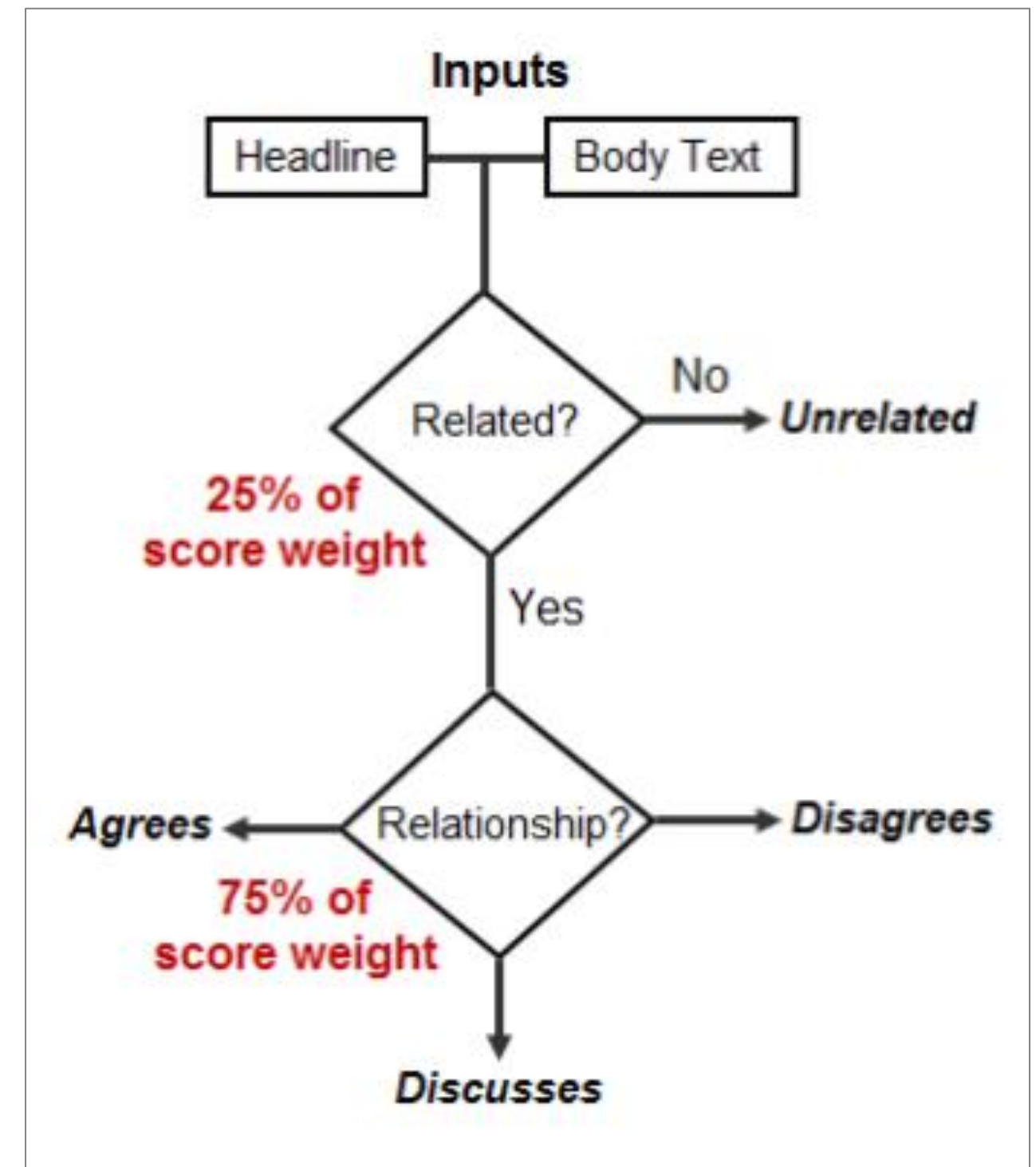
- **Stage 3: Agree/Disagree classification**

- **SVN classifier with class-wise penalty**
- Sentiment and Linguistic Features:
  - Sentiment: using NLTK on first 10 sentences (same as stage 2)
  - Linguistic: LIWC features (<http://liwc.wpengine.com/>) + refuting features (of FNC-I baseline)  
*16 LIWC features: analytical thinking, clout, authentic, emotional tone, conjugation, negation, comparison words, affective processes, positive emotions, negative emotions, anxiety, anger, sadness, differentiation, affiliation, achieve*

# Stance Detection: Evaluation

# Stance Detection: Evaluation Measures

- FNC-I evaluation measure:
  - weighted two-level scoring method
- Rational:
  - The related/unrelated classification task is expected to be much easier and is less relevant for detecting fake news



# Stance Detection: Evaluation Measures

- **Problem:**

- FNC-I evaluation measure does not consider the highly unbalanced class distribution of **neutral (discuss)**, **agree** and **disagree**!

|       | Neutral          | Agree            | Disagree       |
|-------|------------------|------------------|----------------|
| Train | 8,909            | 3,678            | 840            |
| Test  | 4,464 <b>63%</b> | 1,903 <b>27%</b> | 697 <b>10%</b> |

- Not difficult to separate **related** from **unrelated** (the best systems achieves 0.99)
- A classifier that always predicts **neutral** for the related documents achieves **score = 0.83** (same as the top ranked system!)
  - Important to perform well on the important **agree** and **disagree** classes!
- **Better measures:**
  - Class-wise F1 score (harmonic mean of precision and recall for each class)
  - Macro-averaged F1 score



# Stance Detection: Evaluation Results

| System        | FNC         | $F1^m$      | $F1_{Unrel.}$ | $F1_{Neutral}$ | $F1_{Agree}$ | $F1_{Disagr.}$ | $F1^m_{Agree/Disagr.}$ |
|---------------|-------------|-------------|---------------|----------------|--------------|----------------|------------------------|
| Majority vote | 0.39        | 0.21        | 0.84          | 0.00           | 0.00         | 0.00           | 0.00                   |
| FNC baseline  | 0.75        | 0.45        | 0.96          | 0.69           | 0.15         | 0.02           | 0.09                   |
| SOLAT         | 0.82        | 0.58        | <b>0.99</b>   | 0.76           | <b>0.54</b>  | 0.03           | 0.29                   |
| Athene        | 0.82        | 0.60        | <b>0.99</b>   | <b>0.78</b>    | 0.49         | 0.15           | 0.32                   |
| UCLMR         | 0.82        | 0.58        | <b>0.99</b>   | 0.75           | 0.48         | 0.11           | 0.30                   |
| CombNSE       | <b>0.83</b> | 0.59        | 0.98          | 0.77           | 0.49         | 0.11           | 0.30                   |
| stackLSTM     | 0.82        | 0.61        | <b>0.99</b>   | 0.76           | 0.50         | 0.18           | 0.34                   |
| L3S           | 0.81        | <b>0.62</b> | 0.97          | 0.75           | <b>0.53</b>  | <b>0.23</b>    | <b>0.38</b>            |

# Stance Detection: Evaluation Results

- L3S system: Class-wise performance of the different stages

| Stage    | Class     | Precision (P) | Recall (R) | F1 score |
|----------|-----------|---------------|------------|----------|
| Stage 1  | Unrelated | 0.97          | 0.96       | 0.97     |
|          | Related   | 0.91          | 0.93       | 0.92     |
| Stage 2  | Neutral   | 0.82          | 0.80       | 0.81     |
|          | Stance    | 0.67          | 0.71       | 0.69     |
| Stage 3  | Agree     | 0.79          | 0.75       | 0.77     |
|          | Disagree  | 0.40          | 0.44       | 0.42     |
| Pipeline | Unrelated | 0.97          | 0.96       | 0.97     |
|          | Neutral   | 0.74          | 0.76       | 0.75     |
|          | Agree     | 0.52          | 0.53       | 0.53     |
|          | Disagree  | 0.22          | 0.23       | 0.23     |

# Stance Detection: Evaluation Results

- L3S system: confusion matrix

|           | Agree | Disagree | Neutral | Unrelated |
|-----------|-------|----------|---------|-----------|
| Agree     | 1,006 | 278      | 495     | 124       |
| Disagree  | 237   | 160      | 171     | 129       |
| Neutral   | 555   | 252      | 3,381   | 276       |
| Unrelated | 127   | 31       | 523     | 17,668    |

# Stance Detection: Related Problems

- Stance detection of **claim-relevant articles**
- Stance detection of **ideological debates**
- Stance detection of **context-dependent claims**
- Stance detection of **social media posts**


# Stance detection of claim-relevant articles

Wang, X., Yu, C., Baumgartner, S., & Korn, F. (2018, April). Relevant document discovery for fact-checking articles. In Companion of the The Web Conference 2018 (pp. 525-533).

- Given a **fact-checking** article (with schema.org structured markup):
  1. find claim-relevant documents
  2. determine the stance of a claim-relevant document towards the corresponding claim (contradicting / supporting the claim)
- Finding relevant documents:
  - Candidate generation through
    - **navigation** (by exploiting outgoing links and source articles cited in the fact-checking article)
    - **google search** (by running queries using the title of the fact-checking article and text of the claim, as well as *entity annotations* and *click graph queries*)
  - Relevance classification:
    - Classifier: Gradient boosting
    - Features: content similarity, entity similarity, publication order

# Stance detection of claim-relevant articles

Wang, X., Yu, C., Baumgartner, S., & Korn, F. (2018, April). Relevant document discovery for fact-checking articles. In Companion of the The Web Conference 2018 (pp. 525-533).

- Stance classification of **relevant** documents:
  - Binary classification: contradicting / supporting the claim
  - Classifier: Gradient boosting
  - Features: n-grams by exploiting a pre-built **vocabulary for contradiction** and key text (title, headline, important sentences)
- Findings:
  - Relevance classification: 81% precision, 83% recall
  - Stance classification: 96% precision, 86.5% recall
- Limitations:
  - Discuss (neutral) is not considered in stance classification
  - Independent evaluation of relevance and stance classification
    - However: in a real scenario, errors in relevance classification can affect the performance of stance classification
  - Datasets and code are not provided 



# Stance detection of ideological debates

Hasan, K. S., & Ng, V. (2013). Stance classification of ideological debates: Data, models, features, and constraints. In Sixth International Joint Conference on Natural Language Processing (pp. 1348-1356).

- Given a two-sided **debate subject** and an **answer**, determine the **stance of the author** towards the subject
  - Debate subject: *“Should abortion be banned?”*
  - Answer: *“Women who receive abortions are less likely to suffer mental health problems than women denied abortions”*
  - Author stance: Against
- Types of classification models:
  - Independent: assign a stance label to a post independently of the other posts
    - Classifiers: NB, SVM
  - Sequence: consider the linear structure of posts
    - Classifiers: First order hidden Markov model (HMM), linear chain conditional random fields (CRF)
  - Fine grained: jointly determine the stance of a debate post and the stance of its sentences
    - Classifiers: NB, HM

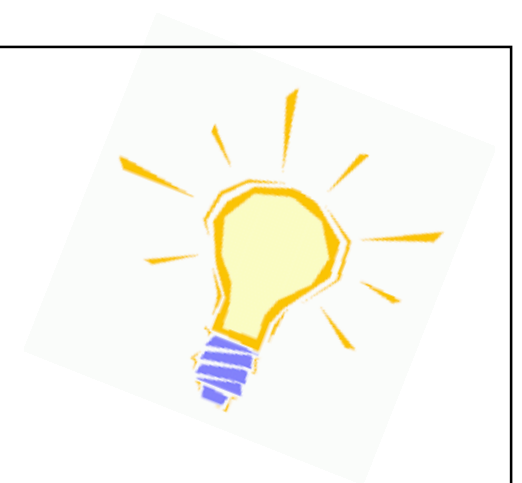
# Stance detection of ideological debates

Hasan, K. S., & Ng, V. (2013). Stance classification of ideological debates: Data, models, features, and constraints. In Sixth International Joint Conference on Natural Language Processing (pp. 1348-1356).

- Features:
  - N-grams (unigrams and bigrams)
  - Document statistics (post length, words per sentence, words with >6 letters, ...)
  - Punctuations (repeated punctuations symbols)
  - Syntactic dependencies (extracting pairs of arguments using a dependency parser)
  - Frame-semantic features (using FrameNet)
- Enforcing author constrains:
  - posts written by the same author for the same topic should have the same stance

## Findings:

- Independent models: no clear winner between **NB** and **SVM**
- **Sequence models** are better than their non-sequence counterparts
- **Fine-grained models** seem to perform better than coarse-grained models
- **Fine-grained HMM** and **CRF** achieve the best results
- **Frame-semantic** features are useful
- **Author constrains** improves the performance of stance classification





# Stance detection of context-dependent claims

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., & Slonim, N. (2017, April). Stance classification of context-dependent claims. In 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1 (pp. 251-261).

- Given a **controversial statement (topic)** and an **argument (claim)**, determine the stance of the argument towards the statement (Pro | Con)
  - Statement/Topic: “Advertising is harmful.”
  - Argument/Claim: “Marketing promotes consumerism and waste.”
  - Stance of claim towards the topic: Pro
- Semantic model:
  - Extract the **target** (e.g., “advertising”) and **sentiment** (1 or -1) of both claim and topic\*
  - Detect the **contrast relation** between the target topic and target claim (1 or -1)
  - $\text{stance}(\text{claim}, \text{topic}) = \text{claim sentiment} \times \text{contrast relation} \times \text{topic sentiment}$   
 $-1 \times 1 \times -1 = 1 \text{ (Pro)}$

\* Topic target and sentiment are considered input

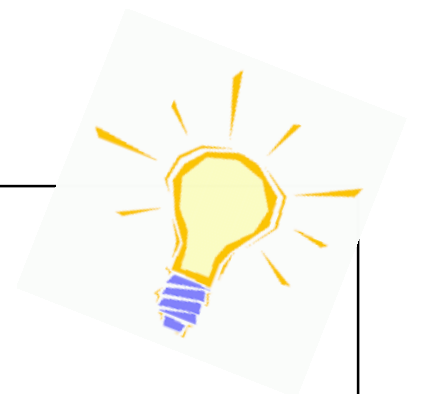
# Stance detection of context-dependent claims

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., & Slonim, N. (2017, April). Stance classification of context-dependent claims. In 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1 (pp. 251-261).

- Claim target identification:
  - Noun phrases in the claim are the candidates
  - Classifier: L2-regularized logistic regression
  - Features: **syntactic and positional** (dependency relation), **Wikipedia** (whether the target is a Wikipedia title), **sentiment** (dependency relation to any sentiment phrase in the claim), **topic relatedness** (semantic similarity between the claim and topic target)
- Claim sentiment classification
  - Lexicon-based sentiment analysis
  - Steps: i) **sentiment matching** (find positive and negative terms), ii) **shifters application** (reverse the polarity of sentiment words), iii) **weighting and scoring** (considering the distance from the claim target)
- Contrast classification algorithm
  - Classifier: random forest
  - Features: contrast scores obtained through relatedness measures

## Findings:

- Accuracy: from 0.63 (considering 100% coverage) to 0.85 (10% coverage)



# Stance detection of social media posts

Du, J., Xu, R., He, Y., & Gui, L. (2017, August). Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

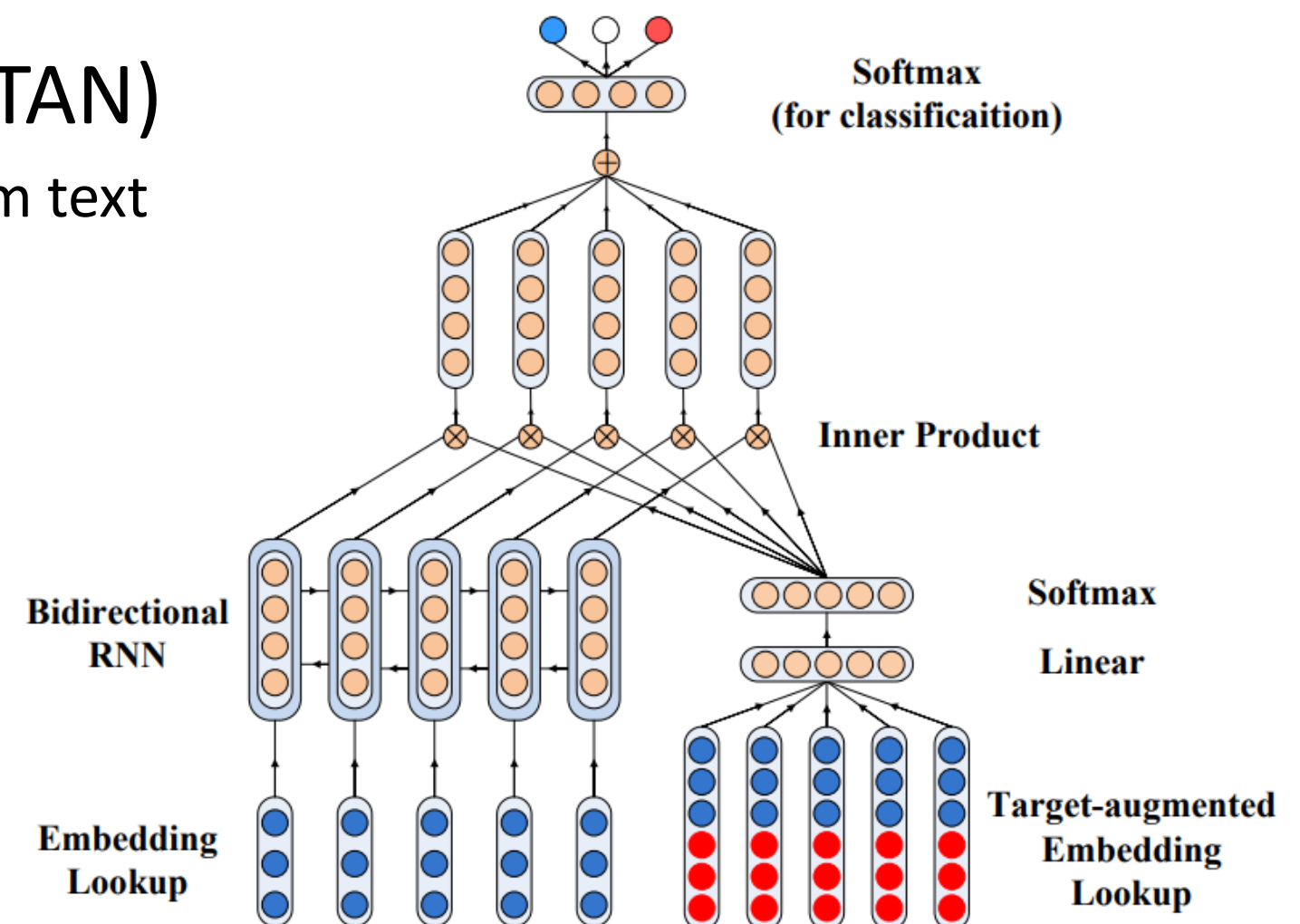
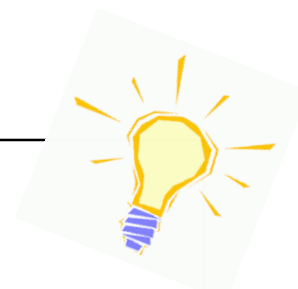
- Given a **controversial topic** and a **post**, determine the stance of the post towards the topic (Pro / Against / NONE)
  - Topic: Abortion
  - Post: “We remind ourselves that love means to be willing to give until it hurts”
  - Stance: Against

## Target-specific Attention Neural Network (TAN)

- RNN with **bi-directional LSTM** for feature extractor from text
- Learn target-augmented embeddings
- Extract target-specific attention signal

### Findings:

- Overall performance in **English** dataset: 69%
- Overall performance in **Chinese** dataset: 73%



# Conclusion

# Concluding remarks



- Fake News
  - Increasingly prevalent nowadays
  - Difficult to detect
  - High impact
- Fact-checking
  - Checking the veracity of a claim
  - Impactful
  - Challenging: laborious, demanding, time-consuming, costly
- Towards facilitating fact-checking:
  - Detecting the **stance** of a document towards a claim

# Concluding remarks




- Stance detection:
  - A supervised (multi-class) classification problem [unrelated, discuss, agree, disagree]
  - Important classes (for fake news): agree / disagree
  - Highly unbalanced class distribution
- Stance detection approaches:
  - Classifiers: gradient boosting, MLP, CNN, SVM with class-wise penalty
  - Features: lexical, sentiment, linguistic, topic model-based
  - Performance: good overall, poor on the important classes (agree / disagree)

Need for models that can better understand the language used to express **agreement** and **disagreement**!

# Thank You!

# Questions?

Pavlos Fafalios

 [fafalios@L3S.de](mailto:fafalios@L3S.de)