

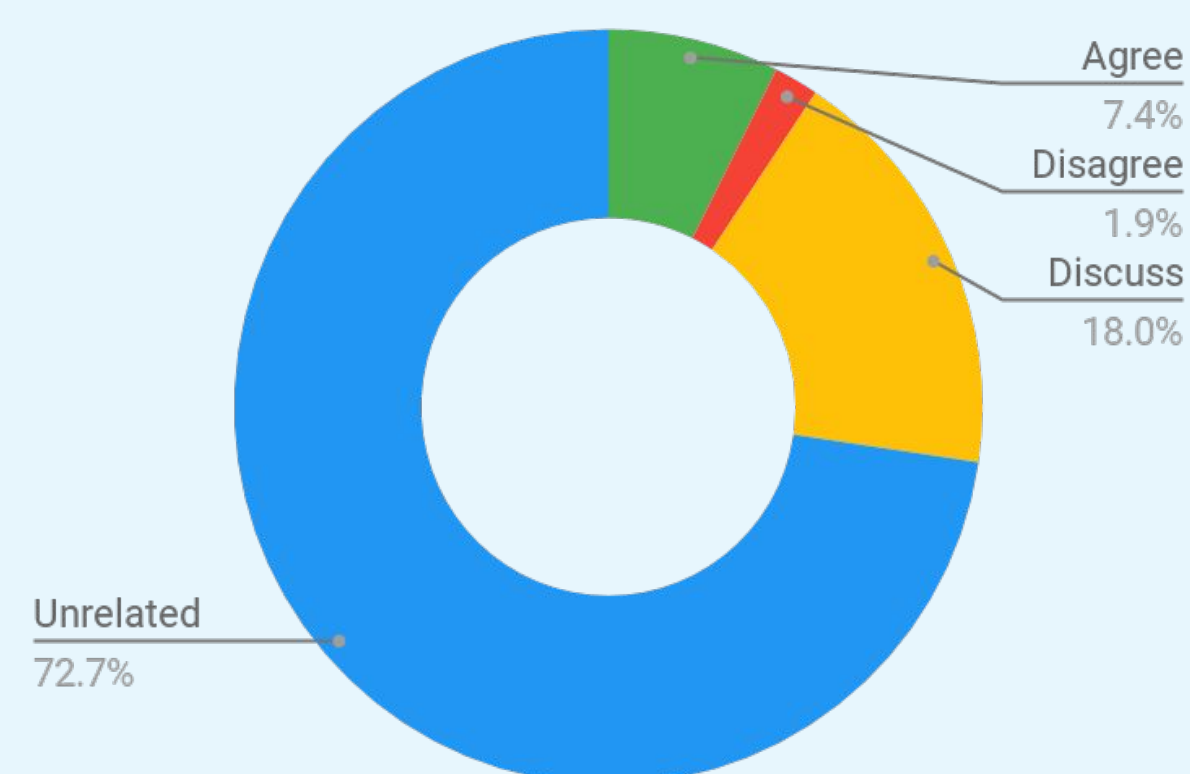
## Problem

- The goal of the Fake News Challenge is to explore how ML and NLP can be used to tackle the Fake news Problem.
- Assessing the correctness of a news story is a complex task, even for trained experts.
- A helpful first step towards **identifying fake news is to understand what other news organizations are saying about the topic**. This process is called Stance Detection.
- It involves Identifying the stance of a news headline, relative to the content of an article which can but does not have to address the same topic.
- The discovery of a disagreeing headline-article pair does not necessarily correspond to the discovery of a fake article, but it is an automated first step which could make human reviewers aware of a discrepancy.

## Data

- The Dataset contains pairs of news headline and article with the appropriate stance label for each.
- Stance can be Agree, Disagree, Discuss, or Unrelated.
- Training dataset is heavily skewed towards Unrelated stance as it is evident from the graph.

Stance Category Distribution in Training Dataset



**Headline** "Robert Plant Ripped up \$800M Led Zeppelin Reunion Contract"

### Body Text Snippets of different Stances

"... Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup. ..." ✓

"... No, Robert Plant did not rip up an \$800 million deal to get Led Zeppelin back together...." ✗

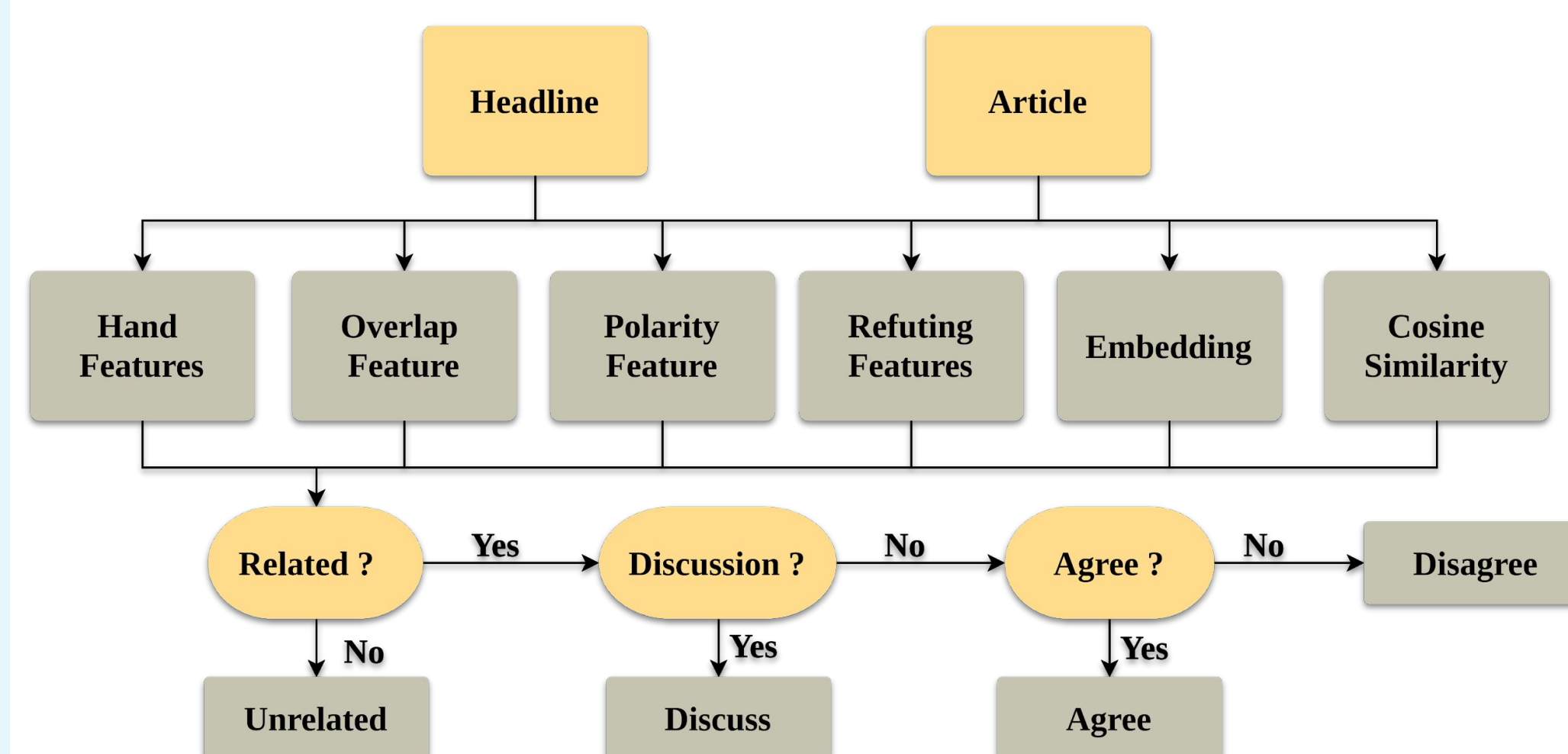
"... Robert Plant reportedly tore up an \$800 million Led Zeppelin reunion deal. ..." ?

"... Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today. ..." !

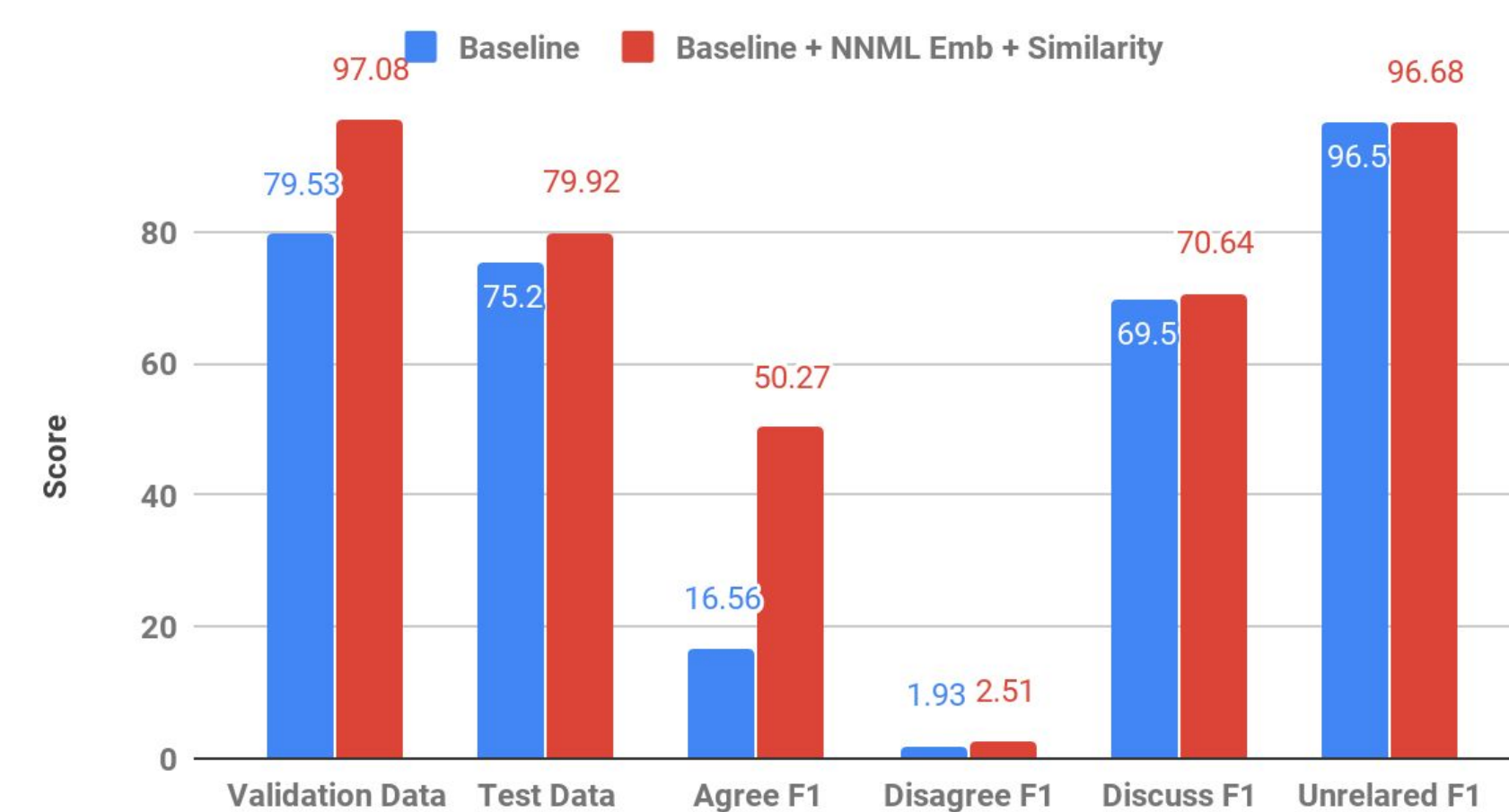
## Approach

- Pre-Processing** of Headline and Article text data
  - Removed the white space and all characters except alphanumeric
  - Converted all the characters to lowercase
  - Tokenized the corpus
  - Applied Lemmatization
  - Removed stopwords only for the hand features
- Features**
  - Overlap Features:** Fraction of overlapped words between headline and body
  - Refuting Features:** Appearance of refuting words in the headline
  - Polarity Features:** Polarity of both headline and body
  - Hand Features:** Counting of occurrence of an n-gram in the headline in the body text
  - Word Embeddings:** Used Pre-trained Neural Net Language Model (NNLM) word embedding of 128 dimensional trained on Google News 200B corpus
  - Cosine Similarity:** Similarity between the Headline and article word embeddings.
- Methods**
  - Word embedding based various configuration of **LSTM/GRU cells**
  - All Features based model with a **Gradient Boosting** algorithm.

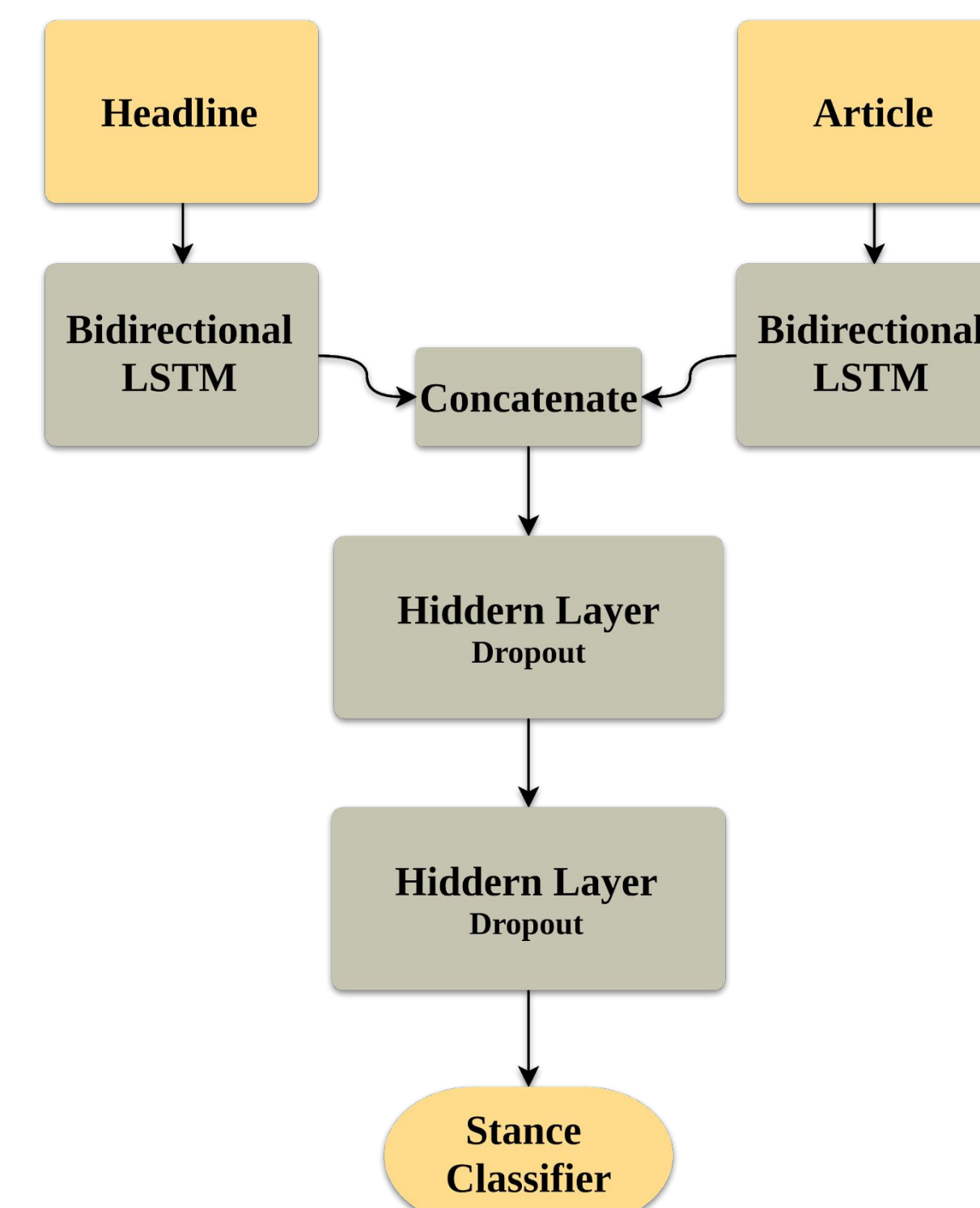
## 3 Step Classification Model



### Baseline and Final Model Comparison



## Bidirectional LSTM Model



## LSTM Model Results

Model	Val Score	Test Score
Simple LSTM	89.28	61.67
LSTM with Attention	91.42	64.14
Bidirectional LSTM	89.75	62.31

## Results for different feature combinations

Features	Val Score	Test Score
Hand	78.5	73.89
Hand + Overlap	79.52	75.12
Hand + Overlap + Polarity	80.33	75.21
3 Step Baseline	80.93	75.67
Embeddings	92.82	66.66
Embedding + Cosine	94.29	73.12
Baseline + Embeddings	96.79	79.06
Final Model	97.08	79.92
Final Model + Stopwords removed	96.69	79.19

## Analysis

- The Developed model increases the score on the test dataset. However, it does poorly on disagreement stance prediction. This could be due to the very few samples of the agree/disagree stance compared to unrelated stance.
- Using the balanced dataset for the agree/disagree classifier (i.e. same article having both agree and disagree stance for different headline) does improve the accuracy of the disagree stance. However, it does at the cost of agree stance's accuracy. Hence, there is no overall improvement.
- Examining the decision tree for the hands only feature:
  - The **related classifier** relies more on how many times a **token of the headline appears in the body text without stopwords**.
  - The **discuss classifier** depends on how many times **bigram of the headline appears in the body text**.
  - The **agree/disagree classifier's** most prominent feature is how many times a **token of the headline appears in the article**.

## Conclusion

- LSTM models with various configuration do quite well on the dev set. However, they fail to beat baseline models on the test set.
- Pre-trained word embeddings perform well without fine-tuning as the dataset is small and fine-tuning the embeddings results in overfitting on the training set.
- More sophisticated feature extraction can be applied to see their effect. Also, Ensemble learning can be used to gain the advantage of pure statistics and deep learning models.
- A balanced dataset can be created to have equal and adequate samples for each stance so that the model can learn to separate stance based on characteristics.

## References

- <http://www.fakenewschallenge.org/>
- Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.
- <https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html>