Modelling Intersecting Context for Fake News Detection

Harsh Mishra, Harsh Bhamore, Aman Sinha and Rohit Agarwal
Indian Institute of Technology (ISM) Dhanbad

Problem

- Shared Task 2: Fake News Detection for the workshop “KDD 2020 True Fact Workshop:Making a Credible Web for Tomorrow” in association with Microsoft.
- Problem Statement :-To detect fake news occuring on social media platforms in the form of articles, posts, etc.
- Processing long texts for having a complete utilization of given dataset to analyze news articles.

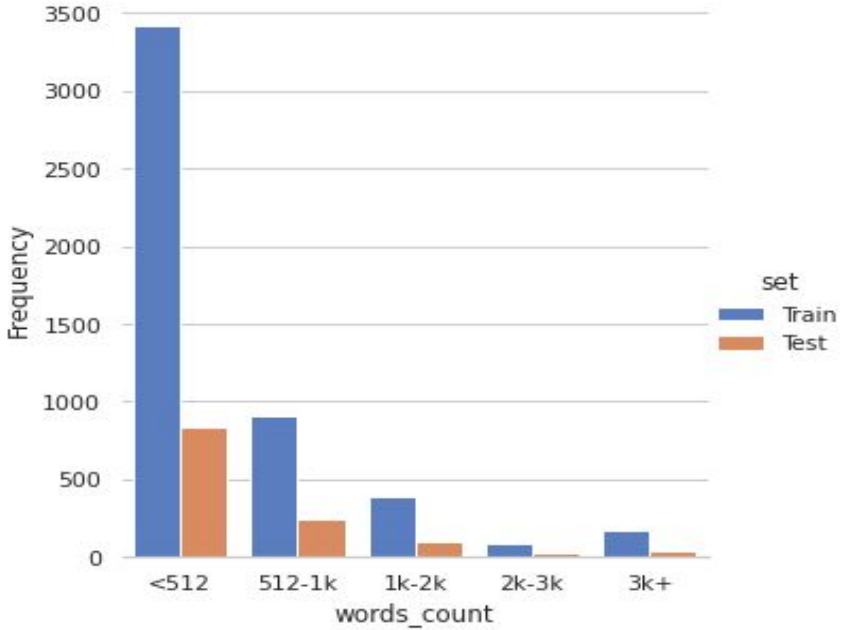
Dataset Instance

"Ruler Elizabeth Says: “Muslim Refugees Are Dividing Nationality, I Fully Agree With Donald Trump We Should Deport Them To Avoid Bloody Terrorist Attacks.” Do you Agree Her?This end of the week the BBC is communicating a journalistic irregularity: A full, sit-down discussion with Queen Elizabeth II.The venture, a review on her royal celebration function in 1953, was 22 years really taking shape, and a media overthrow given the Queen’s memorable hesitance to connect specifically with the press in any capacity.Her Majesty has allowed off camera access to imperial life previously. She likewise gives infrequent broadcast addresses. Be that as it may, “The Coronation,” which affectation on BBC1 at 8pm on Sunday, will be one of her initially broadcast trades with a columnist.It likewise demonstrates her connecting with different crowns engaged with the service, and giving a distinctive portrayal of the experience of being introduced as leader of gigantic swathes of the world (when she took the honored position expansive parts of Africa, the Middle East, and the Caribbean were as yet British settlements)."

Set(Train/Test)	Max.	Min.
Train set	17121 words	4 words
Test set	17520 words	1 words

Full Dataset:- https://www.kaggle.com/c/fakenewskdd2020/data

Challenges



- Long texts are present.
- Considering such a long text, model will face problem of long term dependencies.

Why Modelling Intersecting context ?

"Ruler Elizabeth Says: "Muslim Refugees Are Dividing Nationality, I Fully Agree With Donald Trump We Should Deport Them To Avoid Bloody Terrorist Attacks." Do you Agree Her?This end of the week the BBC is communicating a journalistic irregularity: A full, sit-down discussion with Queen Elizabeth II.The venture, a review on her royal celebration function in 1953, was 22 years really taking shape, and a media overthrow given the Queen's memorable hesitance to connect specifically with the press in any capacity.Her Majesty has allowed off camera access to imperial life previously. She likewise gives infrequent broadcast addresses. Be that as it may, "The Coronation," which affectation on BBC1 at 8pm on Sunday, will be one of her initially broadcast trades with a columnist.It likewise demonstrates her connecting with different crowns engaged with the service, and giving a distinctive portrayal of the experience of being introduced as leader of gigantic swathes of the world (when she took the honored position expansive parts of Africa, the Middle East, and the Caribbean were as yet British settlements)."

Ruler Elizabeth Says: "Muslim Refugees Are Dividing Nationality"

I Fully Agree With Donald Trump We Should Deport Them To Avoid Bloody Terrorist Attacks.

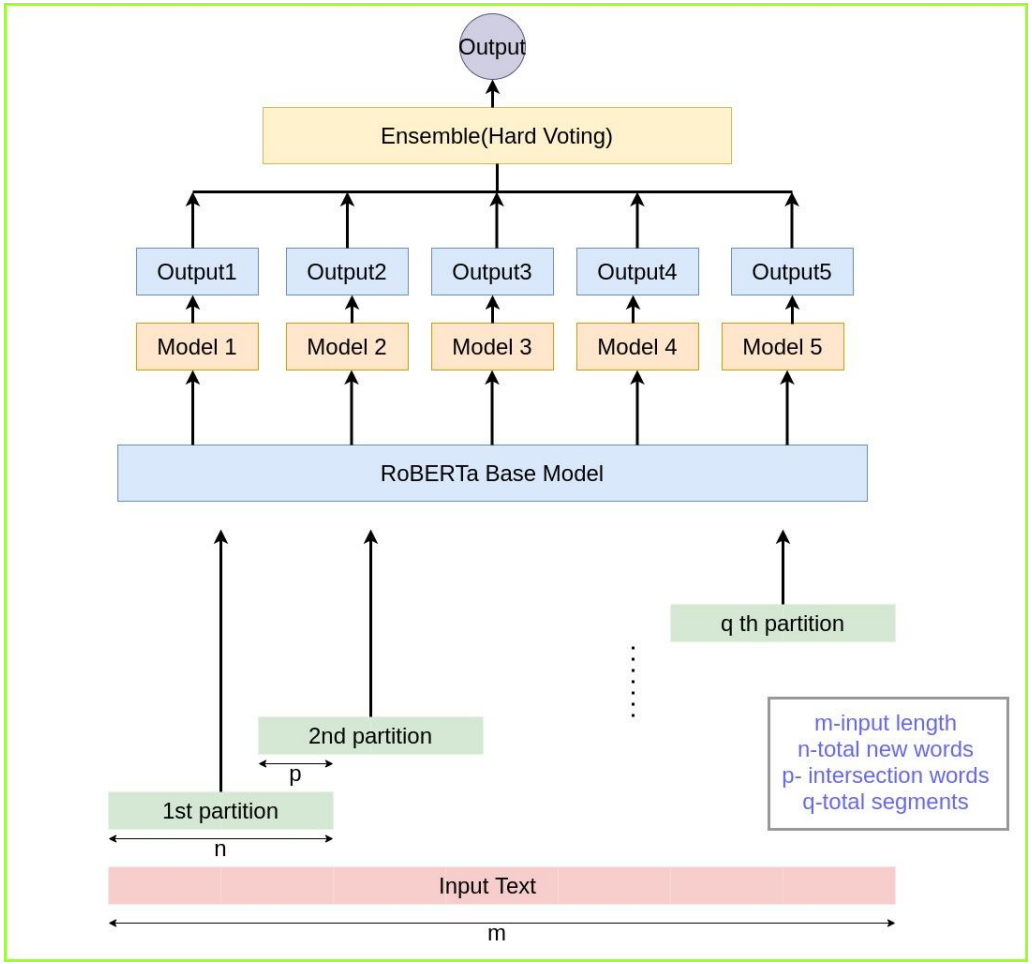
the Middle East,and the Caribbean were as yet British settlements.

- **Annotated label: Fake**
- **Can't use whole long text,let's break it in sub-parts.**

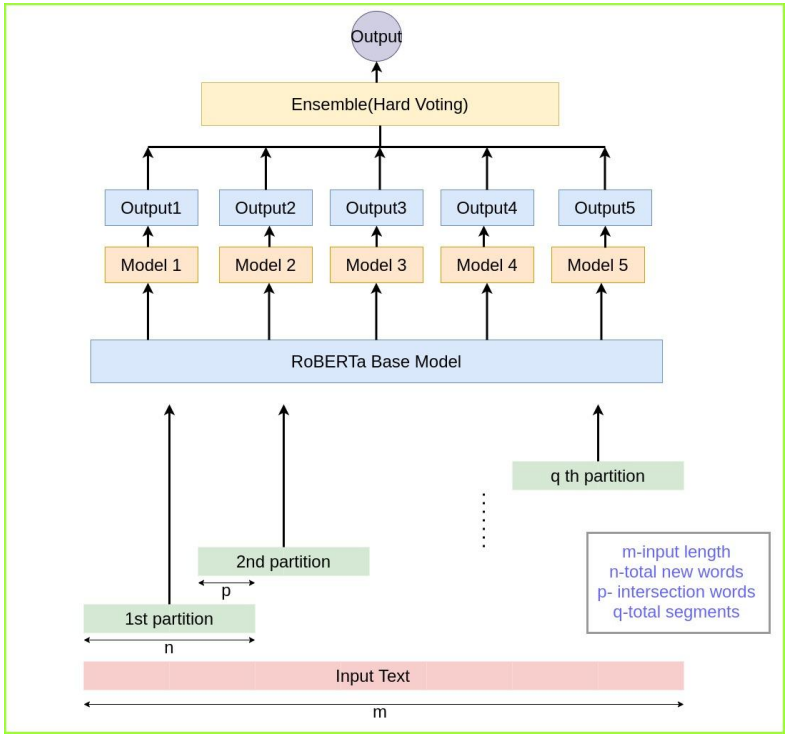
Consider the first two instances,If we label both as fake it would be a problem,The second instance is True,the long text is fake only because of the Previous instance.

Think of a way to add relation between different sub-parts.
"Muslim Refugees Are Dividing Nationality".I Fully Agree With Donald Trump We Should Deport Them To Avoid Bloody Terrorist Attacks.
Hence,The idea of modelling intersection context.

Proposed Model Architecture



Proposed Model



- Long texts partitioned on the basis on intersecting context.
- Fine-tuned k folds on the partitions with label same as long text.
- Ensembled the outputs obtained by different folds(Hard-Voting method).

Experiment Results

Method	Length	Scores
KNN	Full text	0.7470
Naive Bayes	Full text	0.72133
Logistic Regression	Full text	0.78899
SVC	Full text	0.78669
Vanilla GRU	First 300 words	0.6092
Vanilla LSTM	First 300 words	0.52
BERT(base-uncased)	First 128 words	0.7545
XLNet(base)	First 128 words	0.75573
Distil RoBERTa	First 128 words	0.77899
RoBERTa(base-uncased)	First 128 words	0.80504
RoBERTa(base-uncased) with intersecting contexts	n: 150 p: 22	**0.81307**

Discussion

- Hence, RoBERTa model with our approach performed better than baseline RoBERTa. We secured 2nd Rank in the competition.

- Application: We can use this technique for large data analysis for other complex attributes such as humor,sarcasm,etc.

Related Work

- Fake News detection using Naive Bayes((Mykhailo Granik et al.,2017);(Volodymyr Mesyura et al.,2017))
- Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning(Muhammad Syahmi Mokhtar, Yusmadi Yah Jusoh, Novia Admodisastro, Noraini Che Pa, Amru Yusrin Amruddin et al.,2019)
- SVM ((Joshi et al., 2015);(Tepperman et al., 2006); (Kreuz and Caucci,2007); (Tsur et al., 2010); (Davidov et al., 2010)). (Riloff et al., 2013) compared rule-based techniques with an SVM-based classifier
- Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task (Valeriya Slovikovskaya et al.,2019)

Thank you

- Link to our Poster and Code: https://github.com/dsciitism/Fake-News-Detection-Challenge-KDD-2020
- Mail: harshm17172612@gmail.com