Hadith Segmentation Using N-grams

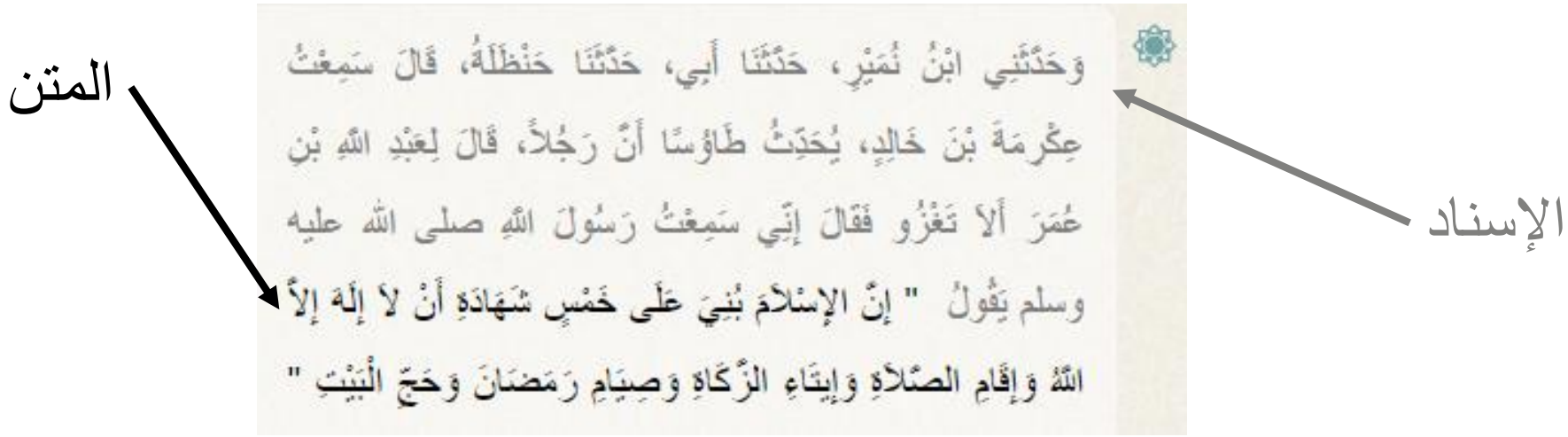
Introduction to NLP (EECE 634) Project Report

Ahmad Mustapha[1], Joseph Sabbagh[2]

[1] Electric and Computer Engineering, AUB

[2]Computer and Communication Engineering, AUB

Context



Sunnah.com

Target: Develop NLP Model to segment Hadith

Training Data

Scraped

Sahih al-Bukhari	Sahih Muslim	Sunan an-Nasa'i	Sunan Abi Dawud
2,910 Hadith	2,999	2,531	1,992

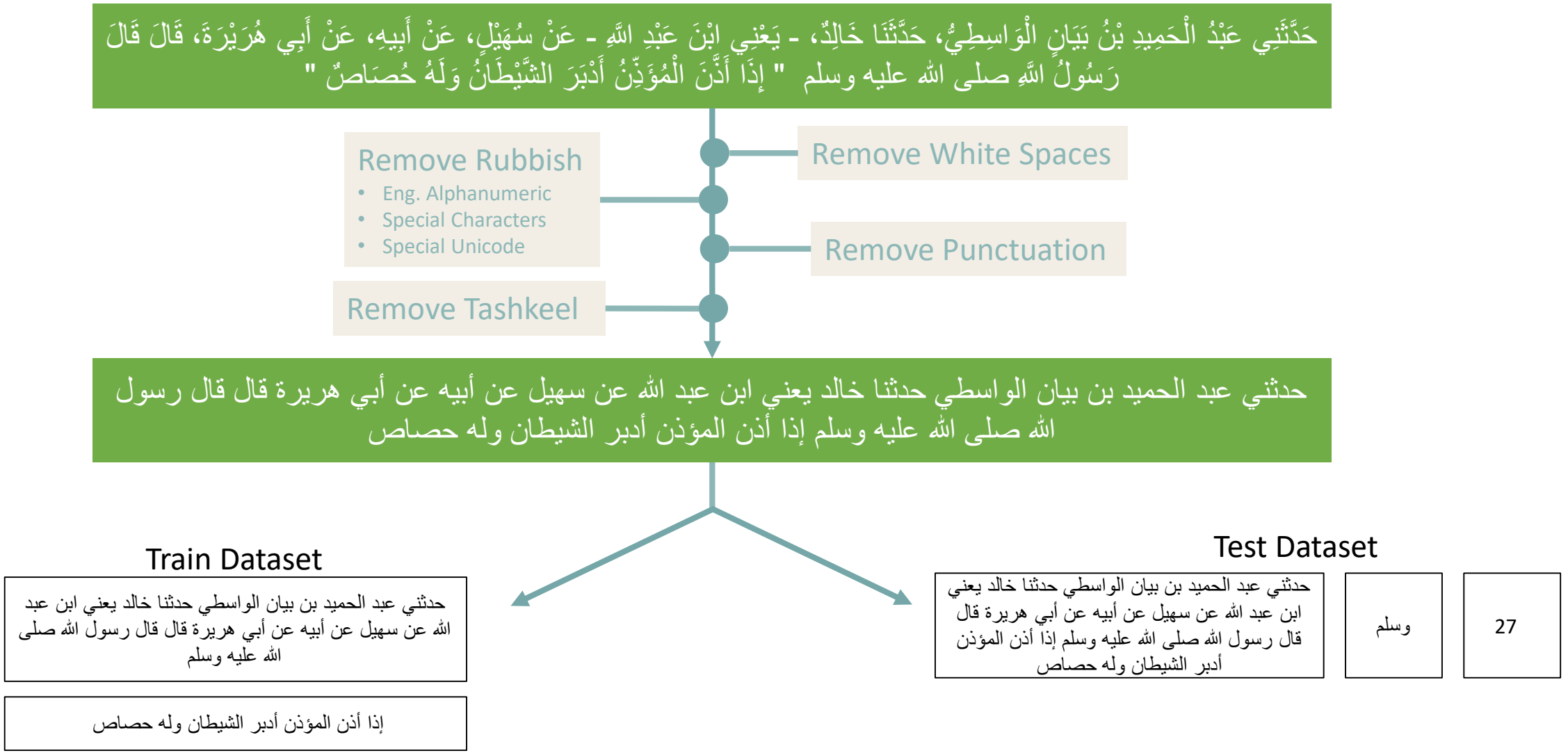
Jami` at-Tirmidhi	Sunan Ibn Majah	Muwatta Malik	Riyad as-Salihin
1,912	2,275	428	933

15,980

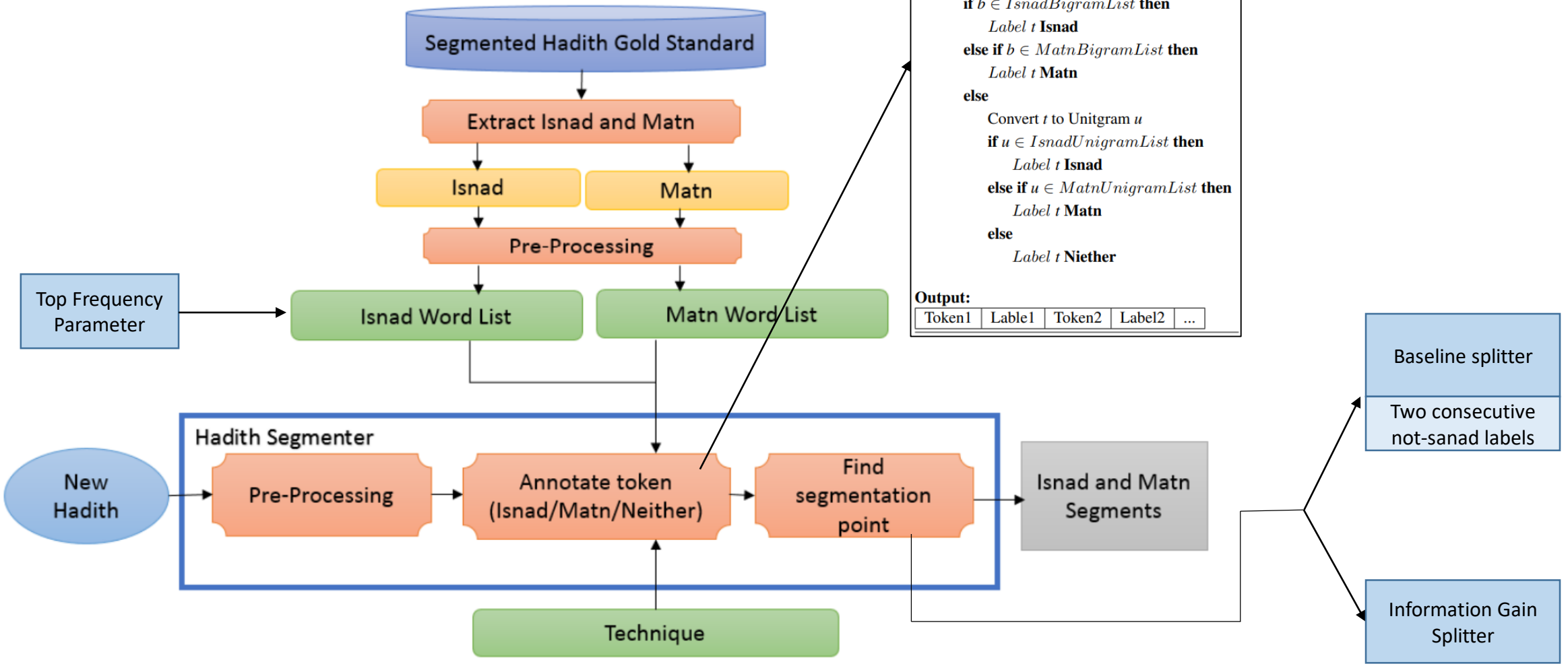
11,186 Train

4,794 Test

Data Processing



Hadith Segmenter



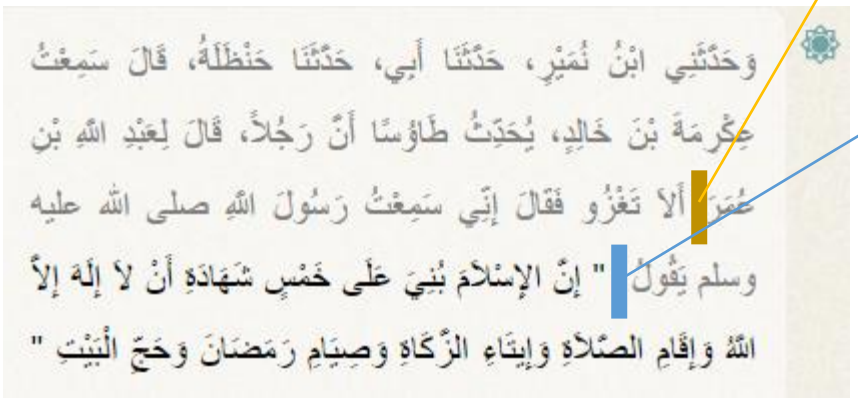
Algorithm 1 Annotate Tri-gram tokens

Tokenize Hadith into Tr-igrams T
for $t \in T$ **do**
  if $t \in IsnadTrigramList$ **then**
    $Label\ t$ **Isnad**
  else if $t \in MatnTrigramList$ **then**
    $Label\ t$ **Matn**
  else
    Convert t to Bigrams b
    if $b \in IsnadBigramList$ **then**
      $Label\ t$ **Isnad**
    else if $b \in MatnBigramList$ **then**
      $Label\ t$ **Matn**
    else
      Convert t to Unitgram u
      if $u \in IsnadUnigramList$ **then**
        $Label\ t$ **Isnad**
      else if $u \in MatnUnigramList$ **then**
        $Label\ t$ **Matn**
      else
        $Label\ t$ **Niether**

Output:

Token1	Lable1	Token2	Label2	...

Evaluation


وَحَدَّثَنِي ابْنُ نُمَيْرٍ، حَدَّثَنَا أَبِي، حَدَّثَنَا حَنْظَلَةُ، قَالَ سَمِعْتُ عِكْرِمَةَ بْنَ خَالِدٍ، يُحَدِّثُ طَاوُسًا أَنَّ رَجُلاً، قَالَ لِعَبْدِ اللهِ بْنِ عُمَرَ أَلاَ تَغْزُو فَقَالَ إِنِّي سَمِعْتُ رَسُولَ اللهِ صلى الله عليه وسلم يَقُولُ " إِنَّ الإِسْلاَمَ بُنِيَ عَلَى خَمْسٍ شَهَادَةِ أَنْ لاَ إِلَهَ إِلاَّ اللهُ وَإِقَامِ الصَّلاَةِ وَإِيتَاءِ الزَّكَاةِ وَصِيَامِ رَمَضَانَ وَحَجِّ الْبَيْتِ "

Predicted Split Position $\widehat{spos}$ — 21

Split Position spos — 33

|33-21| = 12

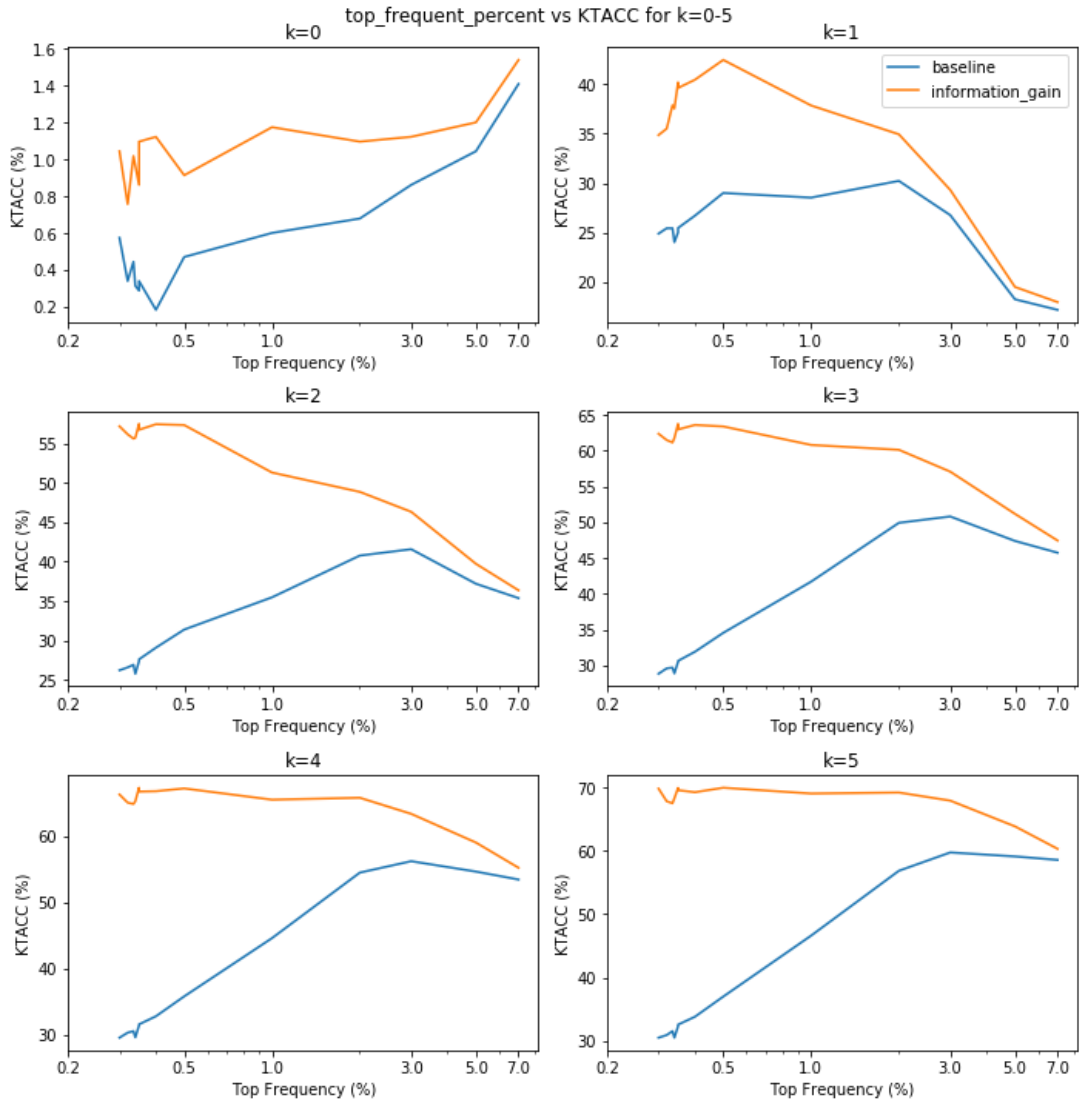
Compare with k

$$Tpred(hadith) = 1(abs(spos - \hat{spos}) <= k)$$

$$Fpred(hadith) = 1 - Tpred(hadith)$$

$$KTACC = \frac{\sum_{h=0}^{H} Tpred(h)}{\sum_{h=0}^{H} Tpred(h) + \sum_{h=0}^{H} Fpred(h)}$$

Fine Tuning


top_frequent_percent vs KTACC for k=0-5
k=0
k=1
k=2
k=3
k=4
k=5
baseline
information_gain
KTACC (%)
Top Frequency (%)
0.2
0.5
1.0
3.0
5.0
7.0

Thank you