Unsupervised Anomaly Detection

Karthik Yadav • ES14BTECH11009
Safwan Mahmood • ES14BTECH11017

Abstract

- Anomaly detection is an imperative issue being addressed by many researchers inside differing research regions and application areas.
- Numerous anomaly detection methods have been particularly created for specific application areas.

- In this paper, we will be evaluating several variations of a technique for finding unusual segments in a document. The methods will be using proven stylistic features to characterize segments of writing.
- Few use cases : Ad-detection, Text translations, Fact vs Opinion detection etc.

What is an Anomaly ?

- Anomalies are patterns in text that don't fit in with a very much characterized idea of ordinary conduct.

- Anomalies may be instigated in the text for an assortment of reasons, for example, pernicious action, e.g., credit card extortion, digital interruption, fear based oppressor action or breakdown of a framework, yet the majority of the reasons has a common feature that is it fascinating to the analyst.

Example ?

Below are few quotes :

- Love all, trust a few, do wrong to none.
- A fool thinks himself to be wise, but a wise man knows himself to be a fool.
- We know what we are, but know not what we may be.
- Some people never go crazy. What truly horrible lives they must lead.
- What's in a name? That which we call a rose by any other name would smell as sweet.

Example ?

Below are few quotes :

- Love all, trust a few, do wrong to none.
- A fool thinks himself to be wise, but a wise man knows himself to be a fool.
- We know what we are, but know not what we may be.
- Some people never go crazy. What truly horrible lives they must lead.
- What's in a name? That which we call a rose by any other name would smell as sweet.

The example had several quotes by William Shakespeare and one by Charles Bukowski.

Scenario :

- A lot of papers have already focused on solving Anomaly detection in a **supervised** environment.
- Here, we would like to approach this in a novel, **unsupervised** manner.
- To ensure efficiency and accuracy, we will be using proven stylistic features to characterize segments of writing.
- We characterise each segment to form vectors which will be used for ranking.

Method 1 : Algorithm in the paper

- Components of our vector representation for a segment consist of simple surface features such as *average word and average sentence length*, the *average number of syllables per word*, together with a range of Readability Measures.

- The Paper uses two kinds of vectors :
  - Feature Vectors
  - Rank Features

Vector 1 : Feature Vectors

- Percentages of words that are articles, prepositions, pronouns, conjunction, punctuation, adjectives, and adverbs.
- The ratio of adjectives to nouns.
- Percentage of sentences that begin with a subordinating or coordinating conjunctions.
- Diversity of POS tri-grams - this measures the diversity in the structure of a text.

Vector 2 : Rank Features

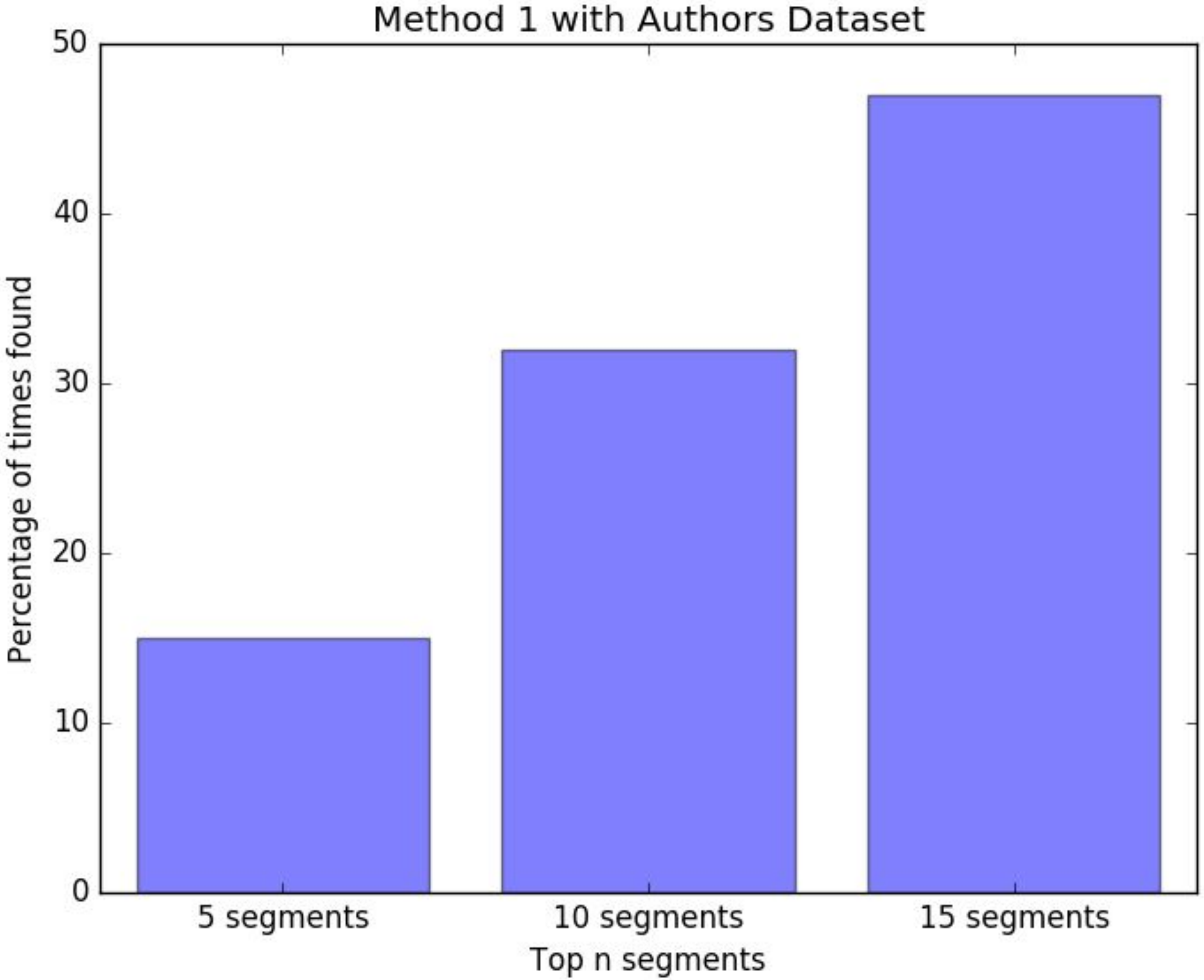
- Most frequent POS tri-grams list
- Most frequent POS bi-gram list
- Most frequent POS list
- Most frequent Articles list
- Most frequent Prepositions list
- Most frequent Conjunctions list
- Most frequent Pronouns list

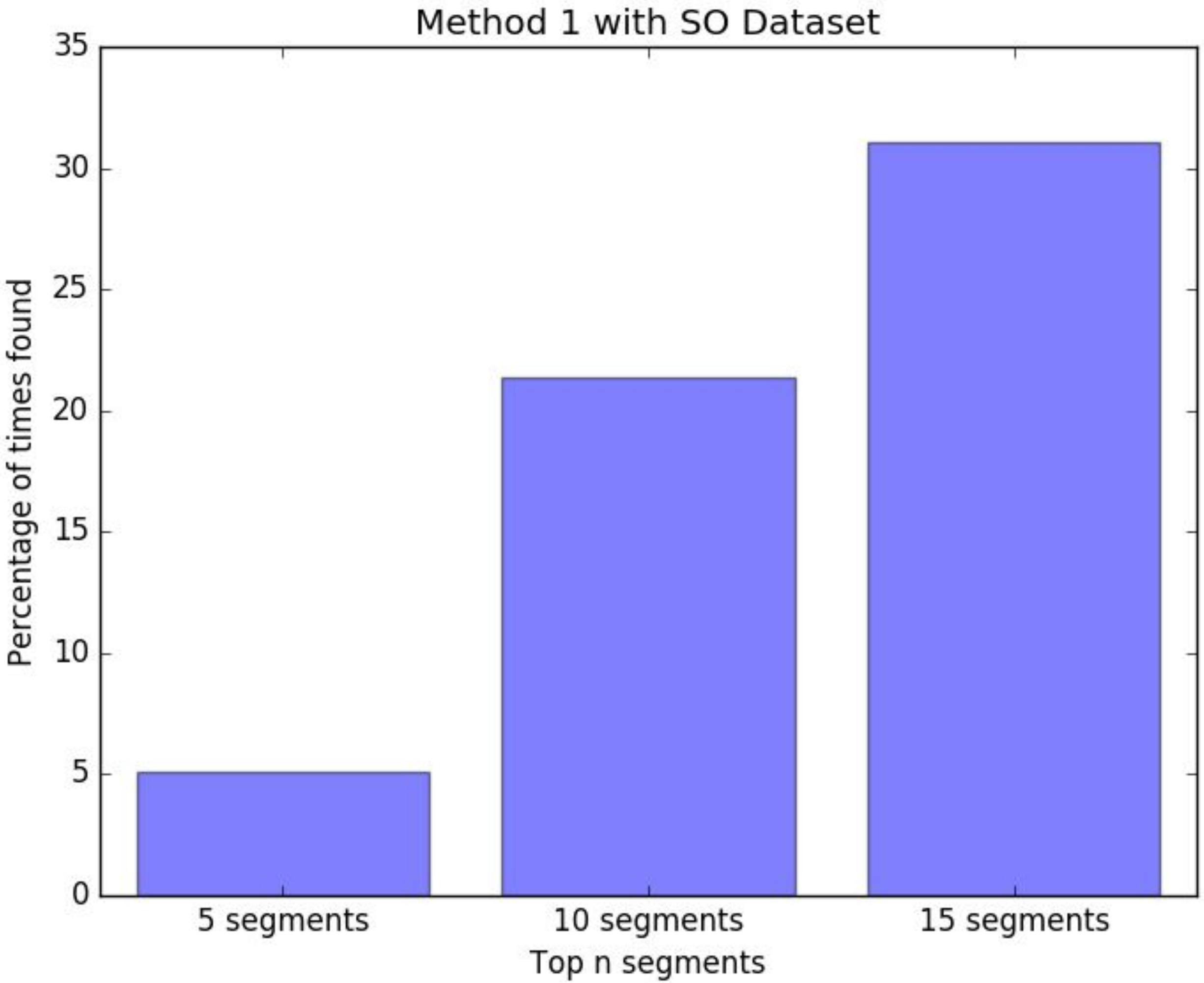
Method 1 : Algorithm in the paper

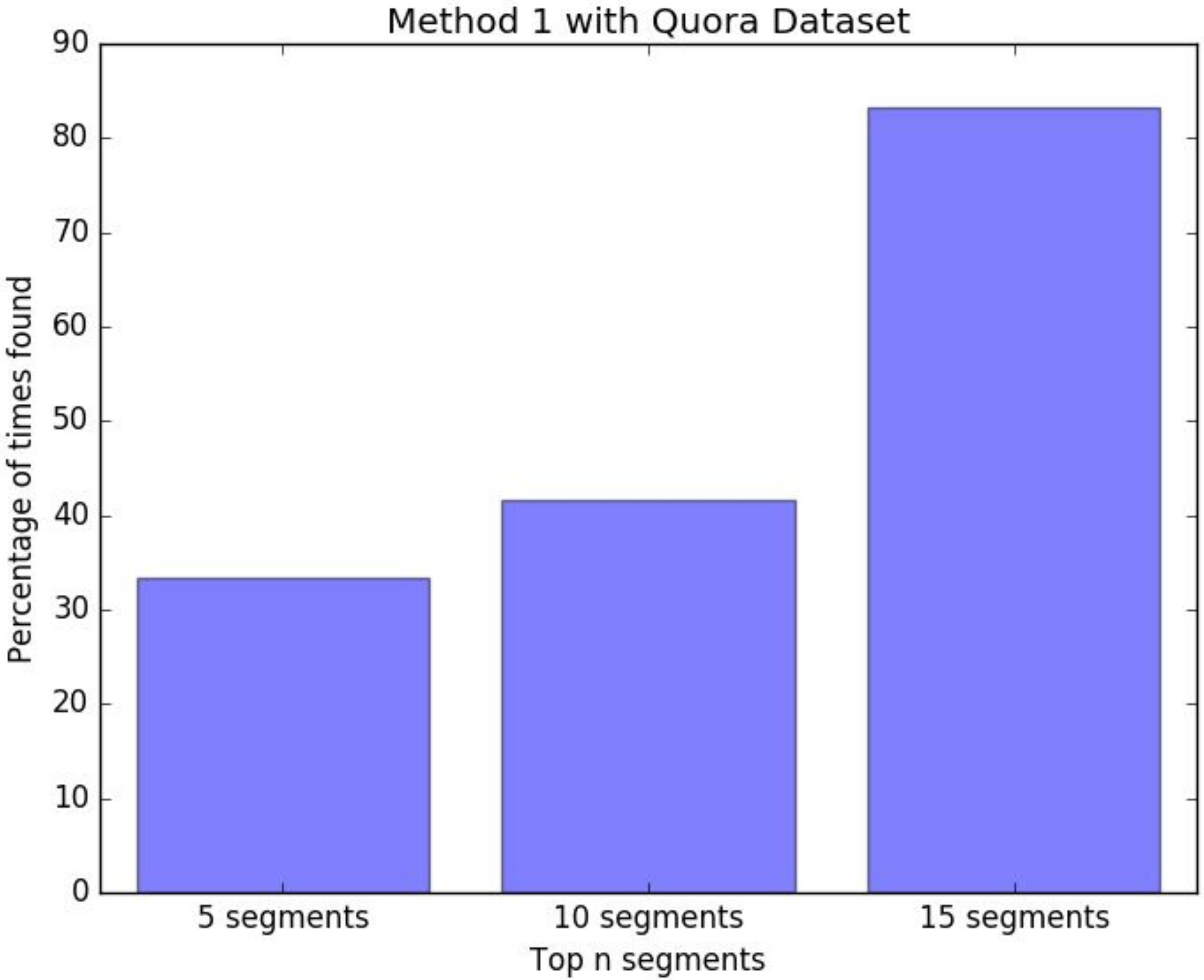
- Assumption : The most anomalous segment of the document is the fake/inserted segment.
- We create these vectors for a particular segment and its complement.
- For features vectors, we take the average difference in their feature vectors (r1).
- For rank features, we use the Spearman Rank Correlation coefficient (r2).
- Using the above information we rank each segment.
- The higher the rank, more the segment is different from rest of the document.

Datasets

1. A list of quotes by various authors : Quotables. Link.
2. StackOverFlow dataset consisting of comments by users on Data Science and Astrology tags. Link.
3. Created a dataset of Quora answers by multiple authors.


Method 1 with Authors Dataset
Percentage of times found
0
10
20
30
40
50
5 segments
10 segments
15 segments
Top n segments


Method 1 with SO Dataset
Percentage of times found
35
30
25
20
15
10
5
0
5 segments
10 segments
15 segments
Top n segments


Method 1 with Quora Dataset
Percentage of times found
0
10
20
30
40
50
60
70
80
90
5 segments
10 segments
15 segments
Top n segments

Challenges :

- We are assuming one segment of the document is anomalous, multiple anomalous segments could be present.
- The writings of a particular author can have multiple styles.
- Method present in the paper doesn’t consider the dependency of semantics in the writings if we consider a general case of StackOverFlow dataset.

Method-2 : Word2vec

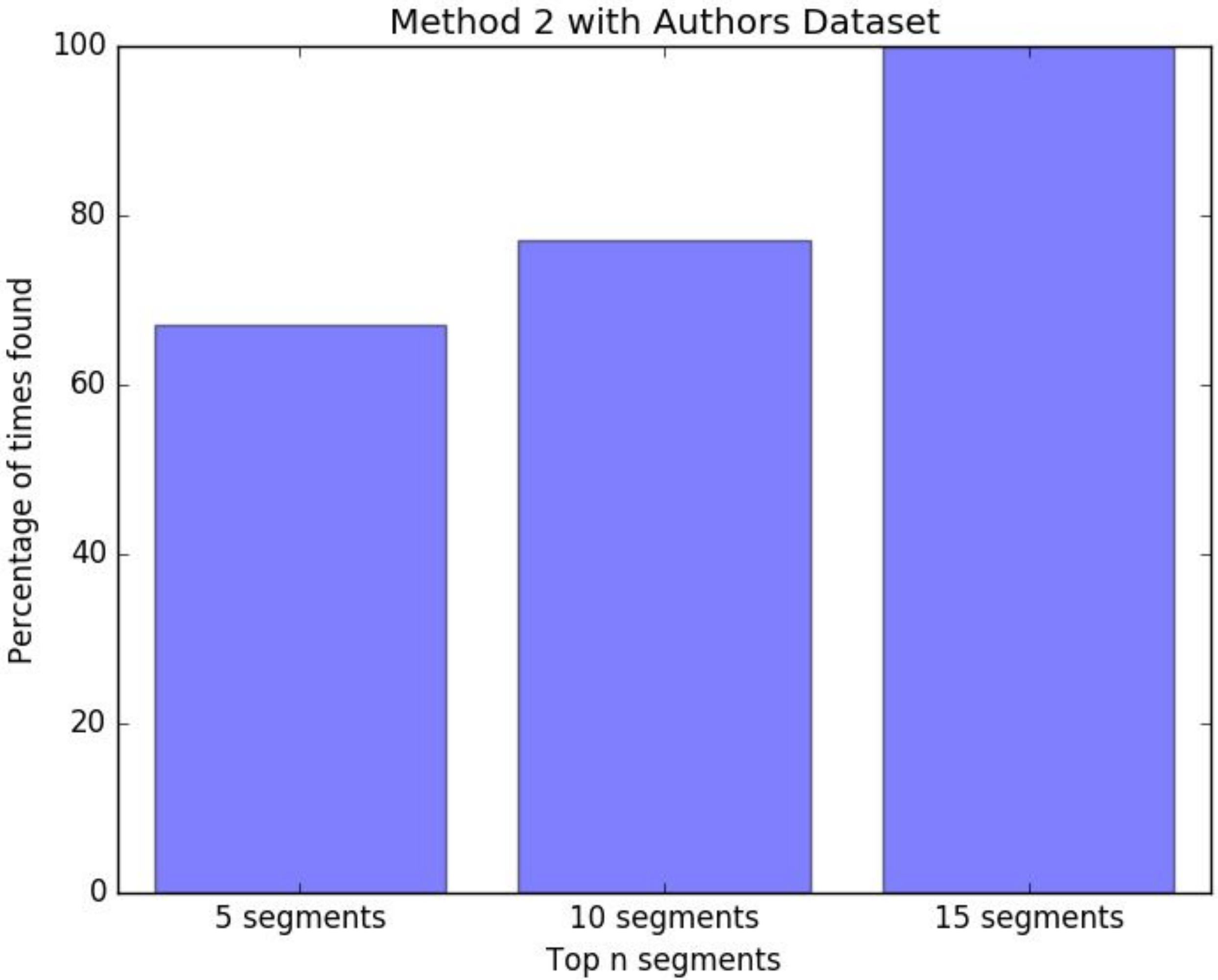
- This is a very standard usage of the Word2vec model.
- We train the Word2vec model with the data (including the anomaly).
- Now, that we have the trained model, we build a similarity matrix for the segments.

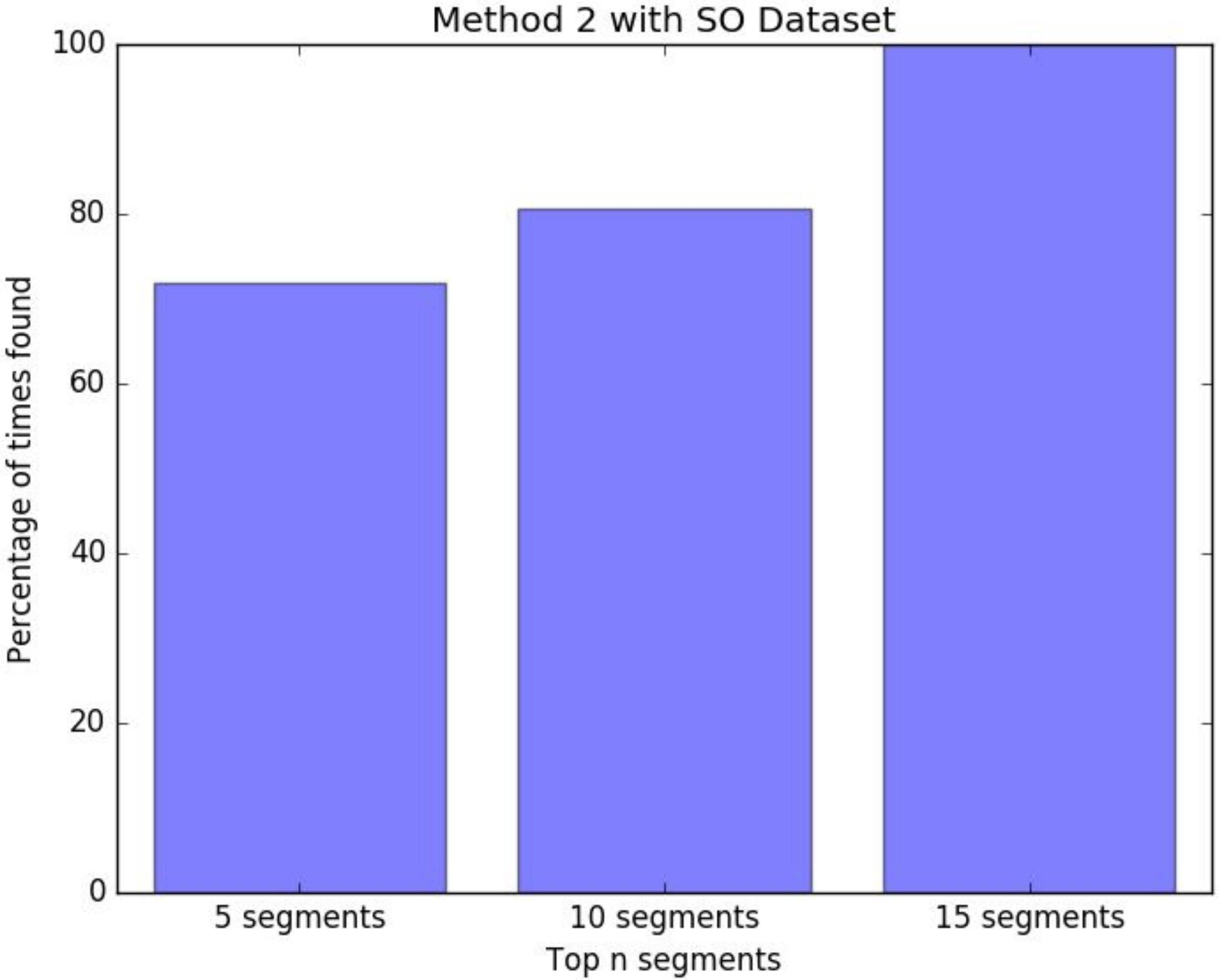
Method-2 : Word2vec

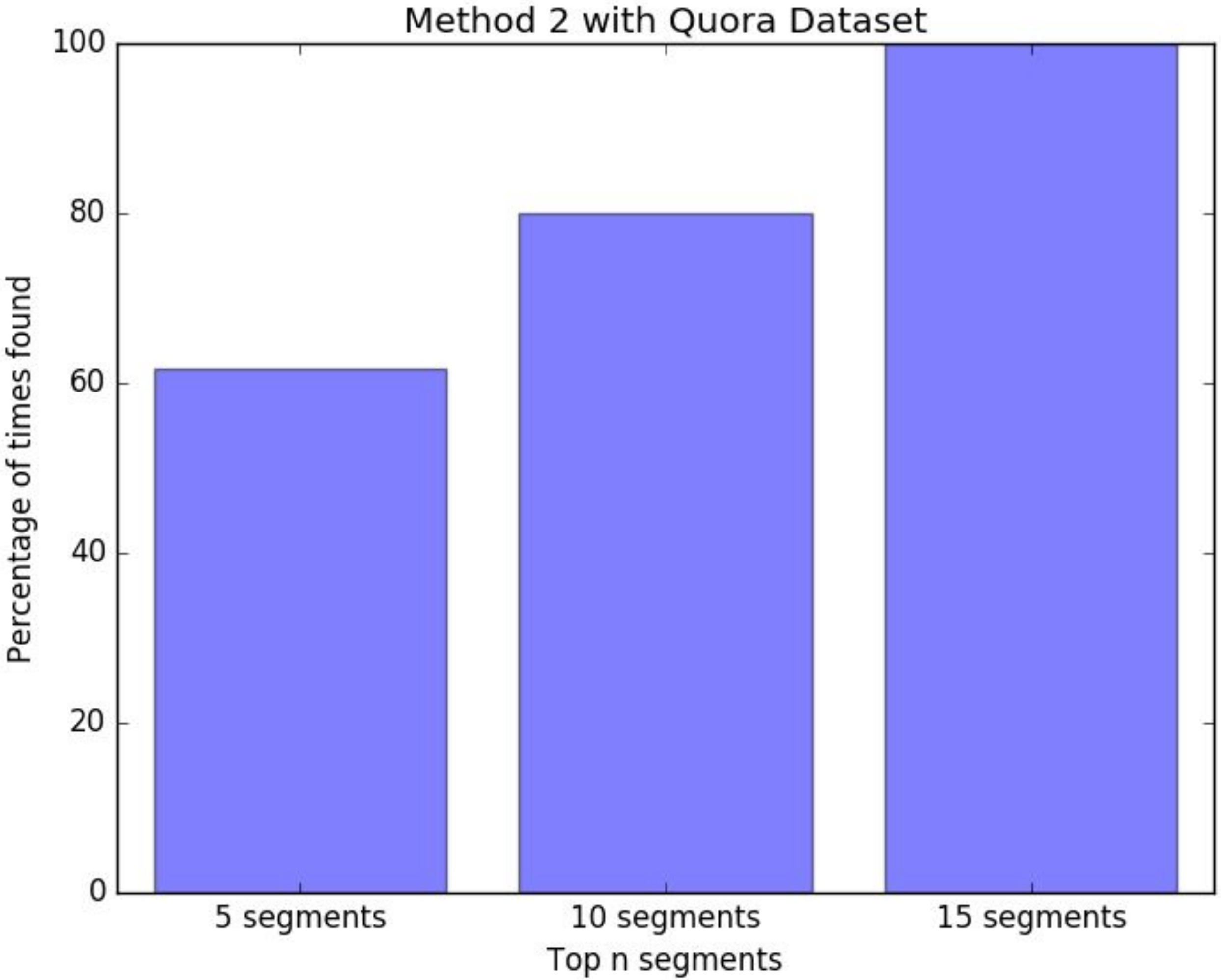
The final ranking score for the i^{th} segment is calculated by:

$$r3_i = \Sigma_{j= 0 \text{ to } n} S_{ij}$$

Lower the ranking score , more different the segment is.


Method 2 with Authors Dataset
Percentage of times found
0
20
40
60
80
100
5 segments
10 segments
15 segments
Top n segments


Method 2 with SO Dataset
Percentage of times found
100
80
60
40
20
0
5 segments
10 segments
15 segments
Top n segments

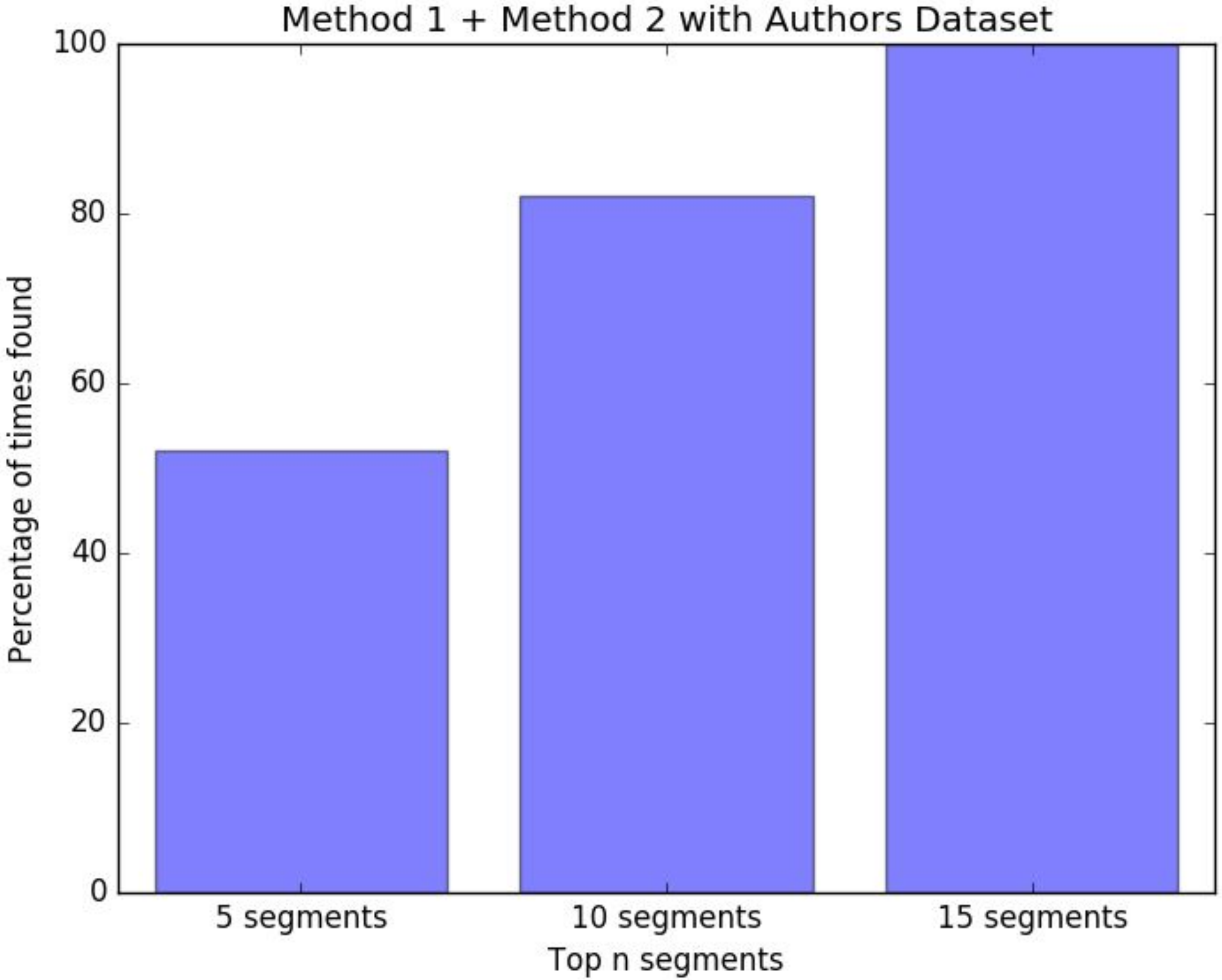

Method 2 with Quora Dataset
Percentage of times found
0
20
40
60
80
100
5 segments
10 segments
15 segments
Top n segments

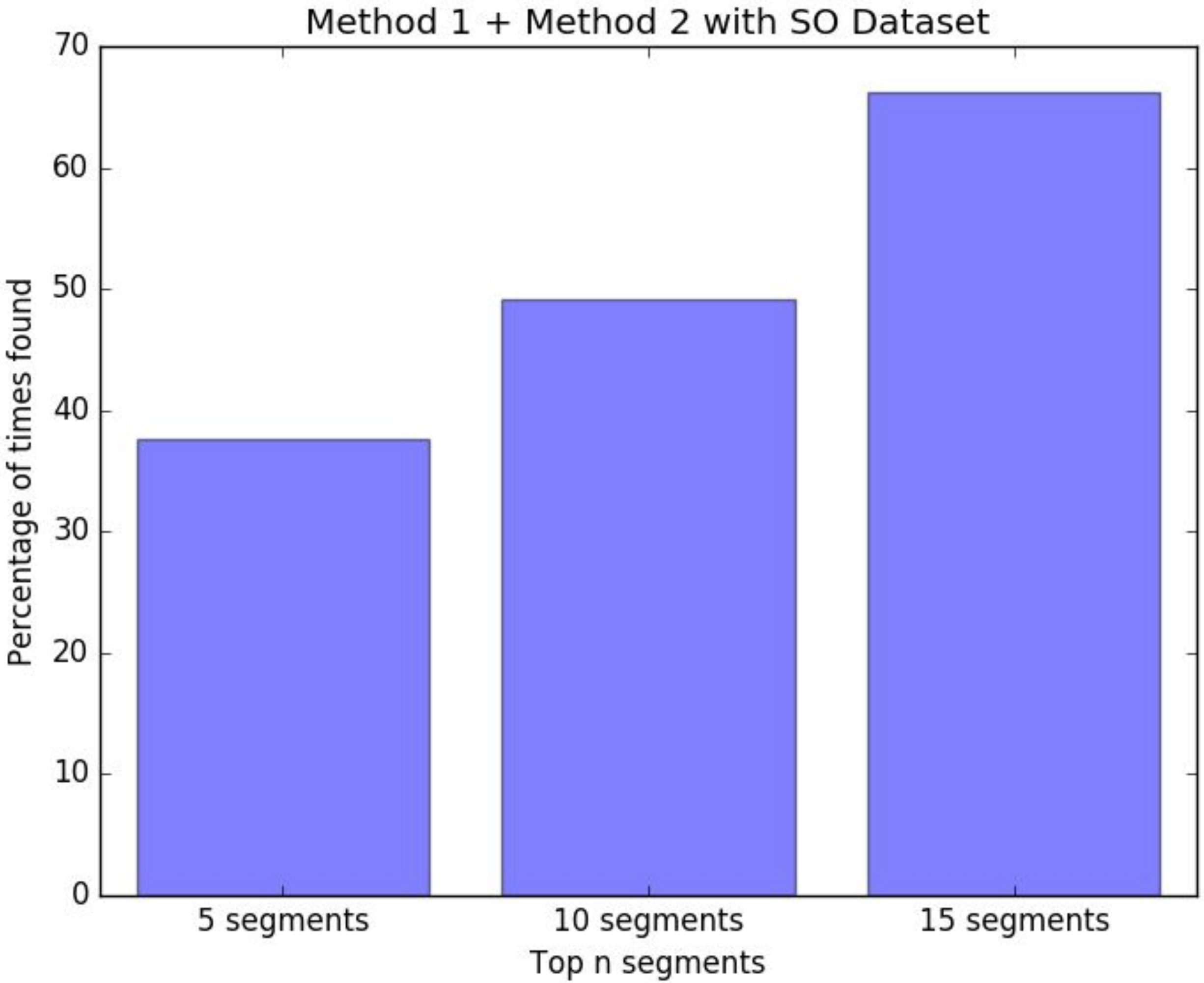
Imbibing semantics into the former code

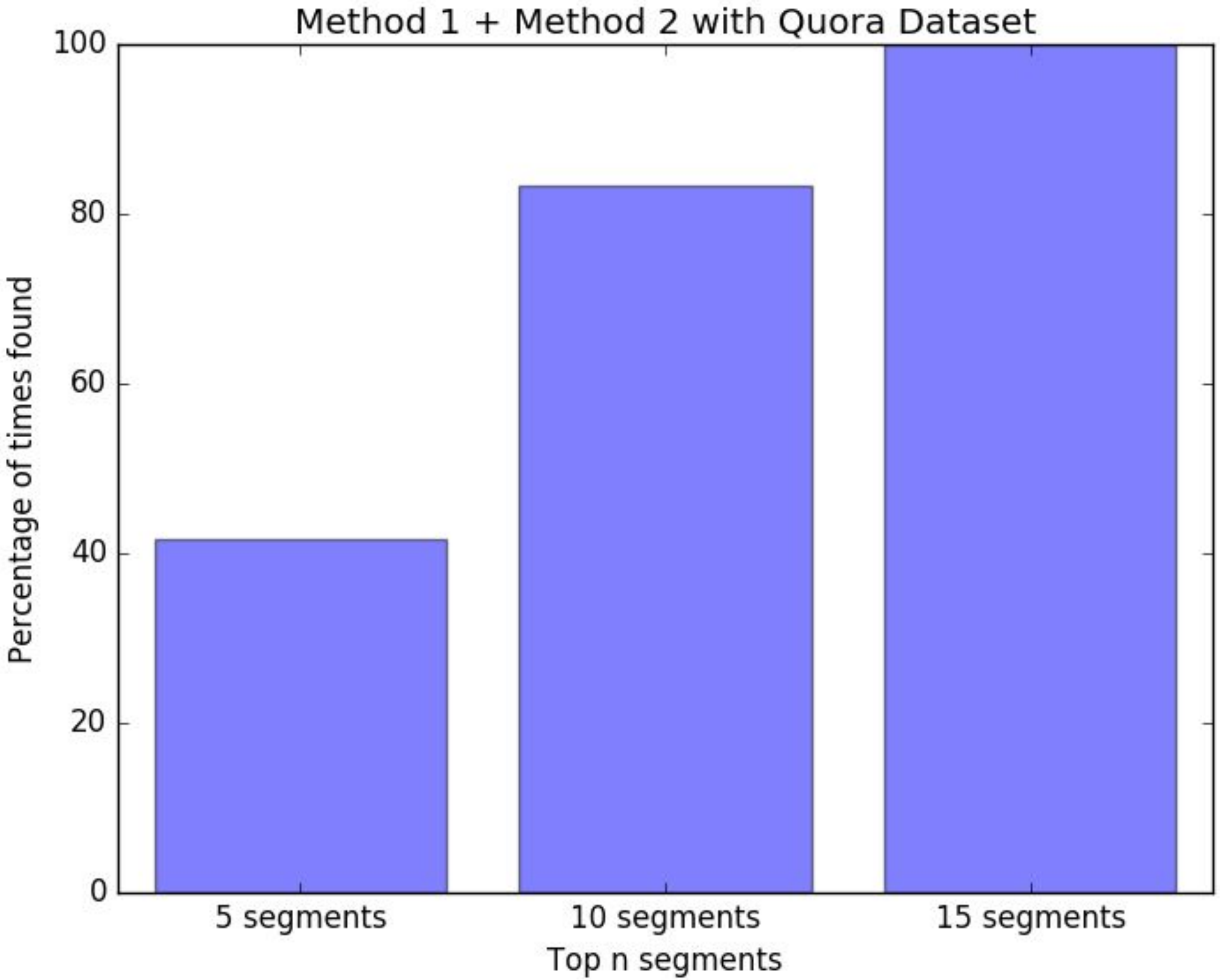
- As we have the ranking scores derived by each method, we combine them to get the final ranking score.

$$R = 0.1 * r1 + 3 * r2 + 2 * (1 - r3)$$

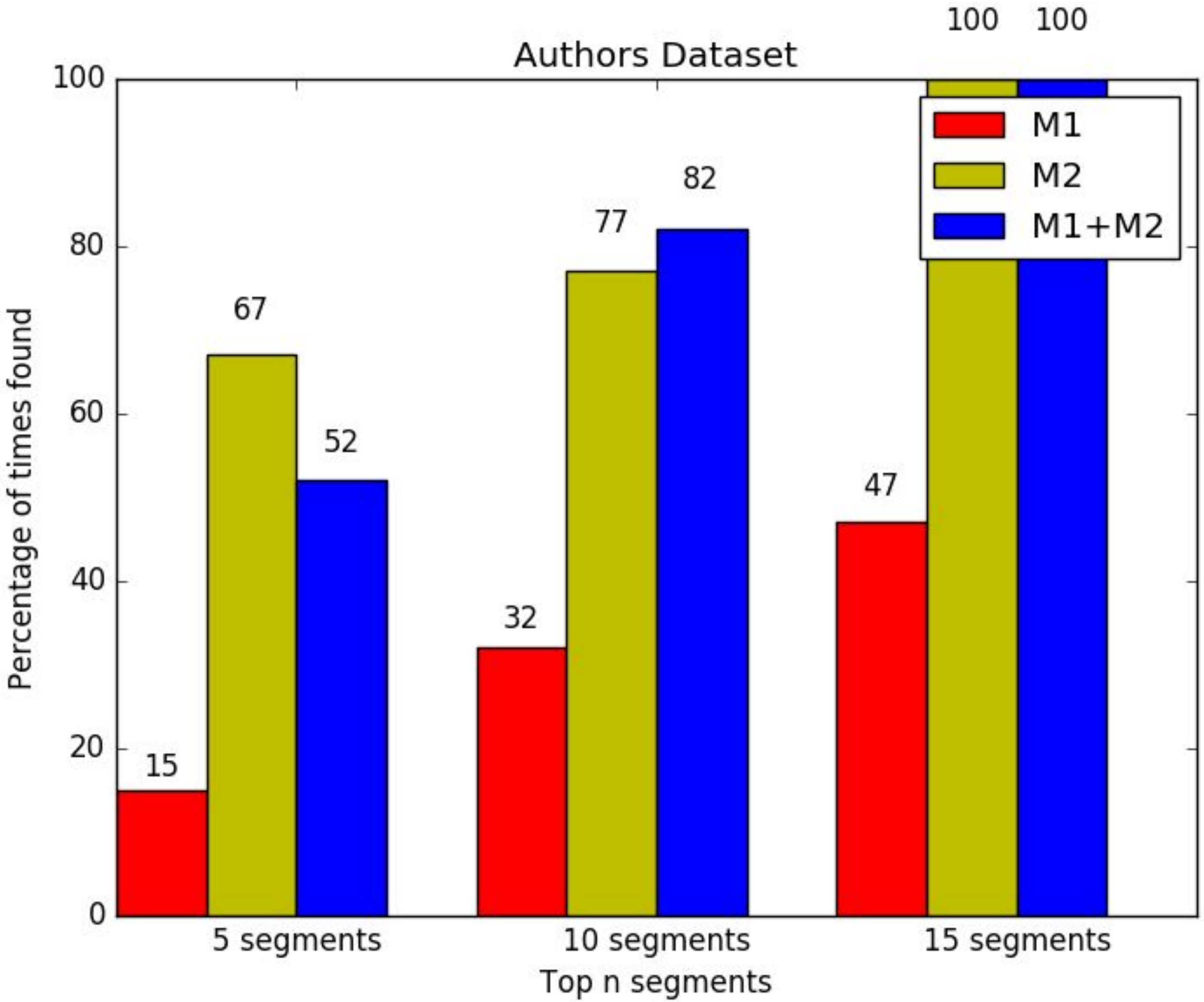
Here, r1 , r2 and r3 are used after normalization.

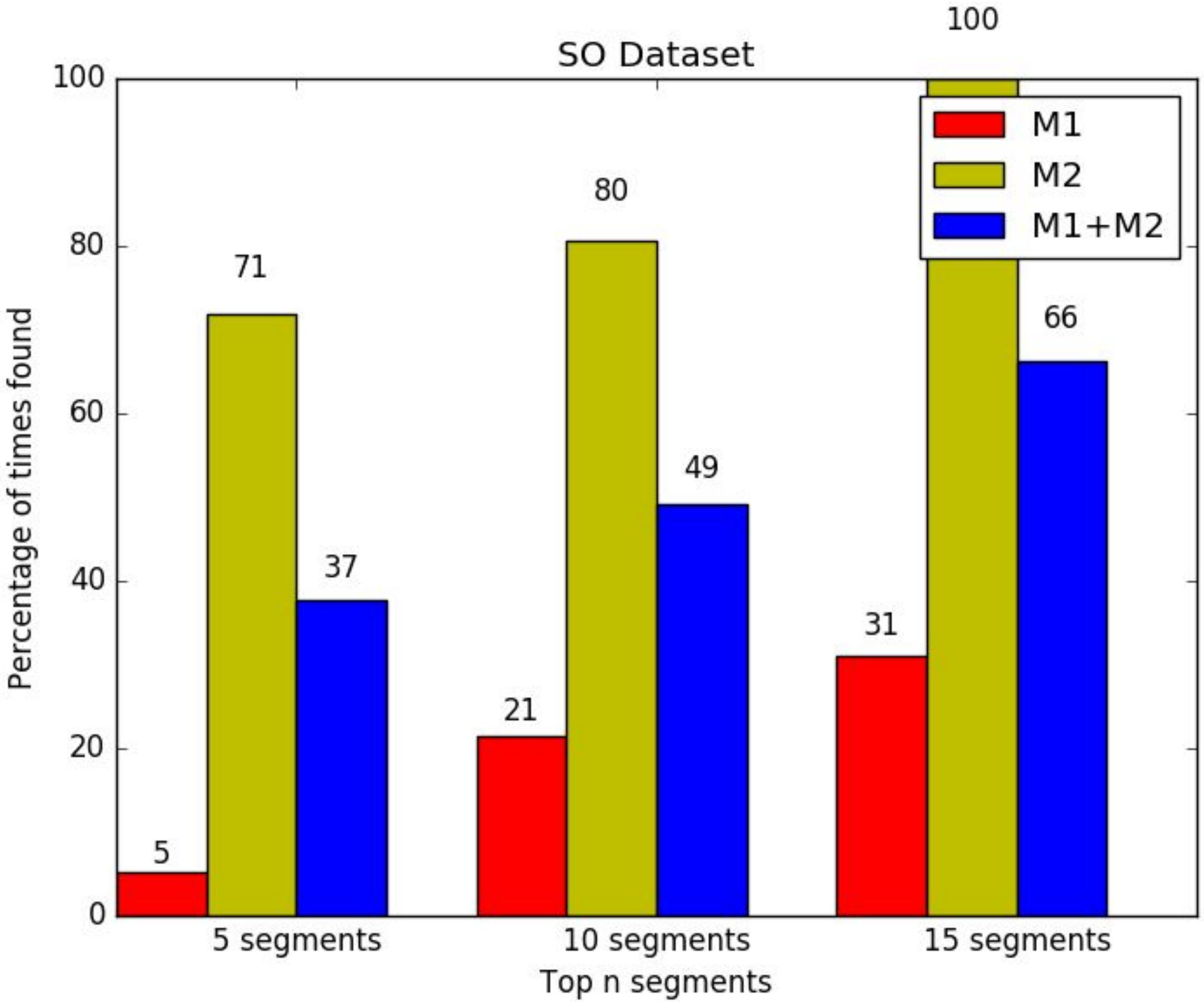

Method 1 + Method 2 with Authors Dataset
100
80
60
40
20
0
Percentage of times found
5 segments
10 segments
15 segments
Top n segments

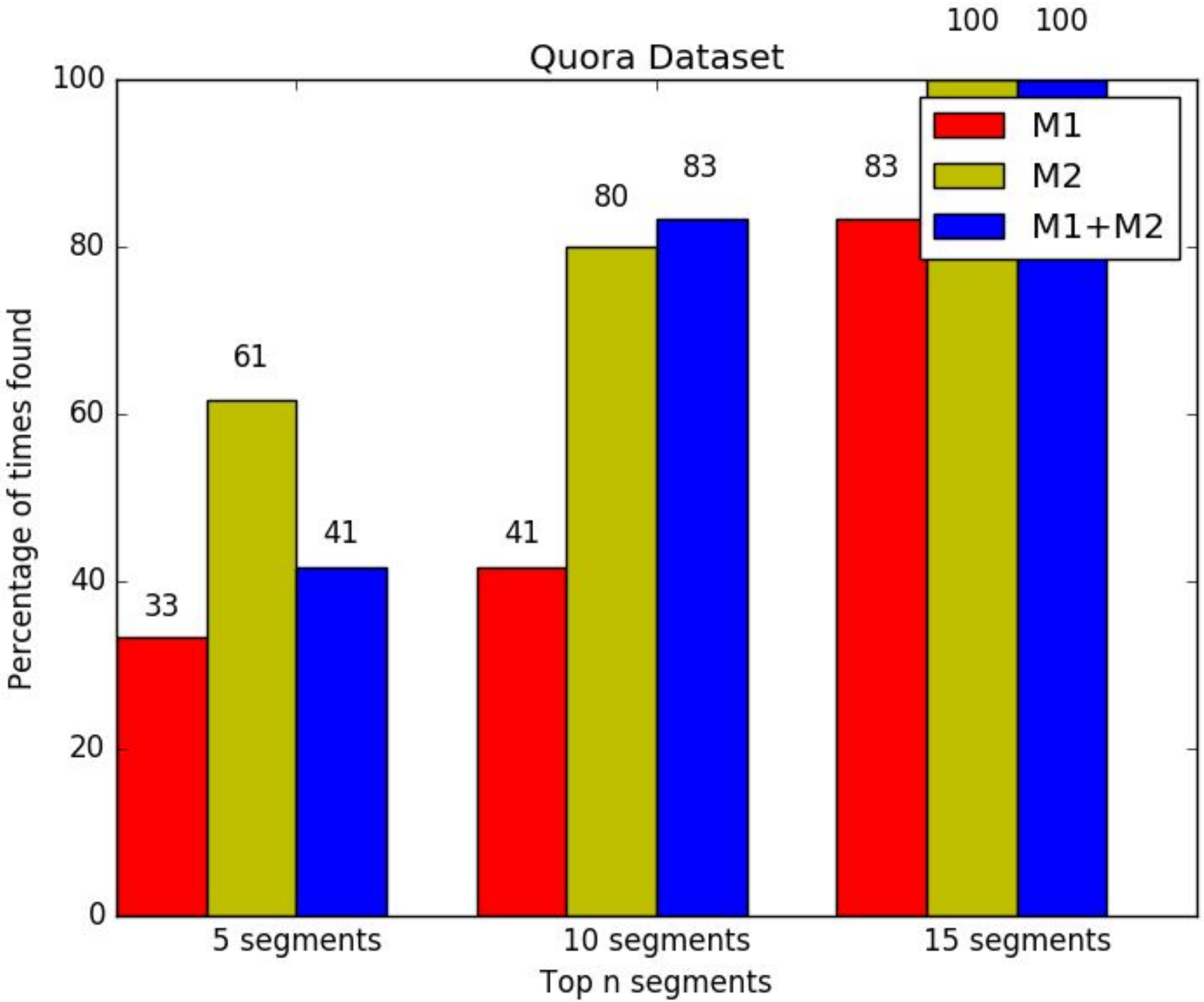

Method 1 + Method 2 with SO Dataset
Percentage of times found
0
10
20
30
40
50
60
70
5 segments
10 segments
15 segments
Top n segments


Method 1 + Method 2 with Quora Dataset
Percentage of times found
0
20
40
60
80
100
5 segments
10 segments
15 segments
Top n segments

—


Authors Dataset
100
80
60
40
20
0
Percentage of times found
15
67
52
32
77
82
47
100
100
5 segments
10 segments
15 segments
Top n segments
M1
M2
M1+M2


SO Dataset
M1
M2
M1+M2
Percentage of times found
0
20
40
60
80
100
5
71
37
21
80
49
31
100
66
5 segments
10 segments
15 segments
Top n segments


Quora Dataset
Percentage of times found
0
20
40
60
80
100
Top n segments
5 segments
10 segments
15 segments
M1
M2
M1+M2
33
61
41
41
80
83
83
100
100

Conclusion

- The paper considers the style of the writing , ignoring the semantics.
- Imbibing the semantics using Word2Vec always show better results than just the style.
- Word2Vec outperforms the method discussed in the paper, in every case.
- However, the amalgamation of both the methods outperforms both individual methods in most cases.