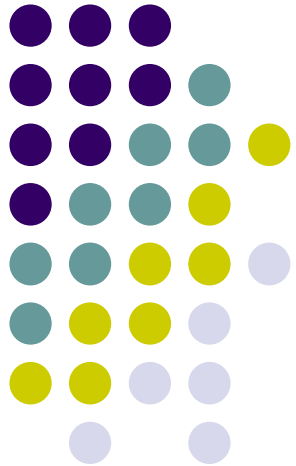
12th International Symposium on Foundations of Information and Knowledge Systems, Helsinki, June 2022

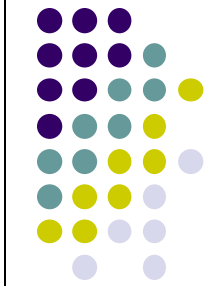
Text classification using "imposter" projections method



Zeev Volkovich

Braude College, Department of Software Engineering
P.O. Box: 78, Karmiel, 21982, Israel
E-mail: vlvolkov@braude.ac.il

Authorship Recognition Problem



Problem

Identification of the authors of an anonymously written set of documents given a set of candidate authors.

Applications:

- Determining issues of doubtful authorship
- Identifying authorship of anonymous texts
- Detecting plagiarism
- Investigative Forensic Identification
- E-mails, tweets, posts
- Identifying Pseudepigraphic.

Some Classical Examples

1. Did Moses Write the Pentateuch?
2. Who Wrote the Analects?
3. Who Wrote the Secret Gospel of Mark?
4. Did Abelard and Heloise Write the Letters Attributed to Them?
5. Who Wrote the Compendium of Chronicles (Jami al-Tawarik) and the Collection of Letters Attributed to Rashid al-Din?
6. Who Wrote Shakespeare?
7. Who Wrote the Works Attributed to Prince Andrei Kurbskii?
8. How Inauthentic Was James Macpherson's "Translation" of Ossian?
9. Did Mikhail Sholokhov Write The Quiet Don?

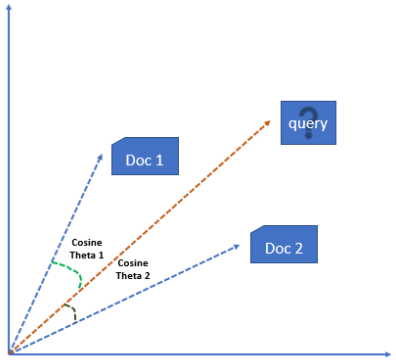
From : Donald Ostrowski "Who Wrote That?"

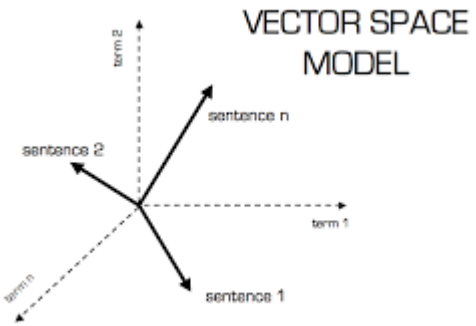
Authorship Controversies from Moses to Sholokhov

NORTHERN ILLINOIS UNIVERSITY PRESS 2020

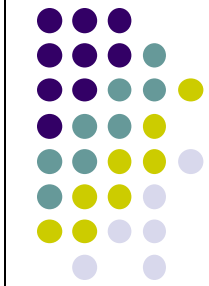
Vector space model

- This is the basic text model, also named term vector model, widely used in the text processing area.
- It is a standard technique in Information Retrieval.
- Allows decisions to be made about which documents are similar to each other and to queries.
- The Bag-of-words model is a partial case of a Vector space model.





‘Bag of words’ representation of text



ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS

BUENOS AIRES, Feb 26

Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, showing those for future shipments month, 1986/87 total and 1985/86 total to February 12, 1986, in brackets:

Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total 2,692.4 (4,161.0).

Maize Mar 48.0, total 48.0 (nil).

Sorghum nil (nil)

Oilseed export registrations were:

Sunflowerseed total 15.0 (7.9)

Soybean May 20.0, total 20.0 (nil)

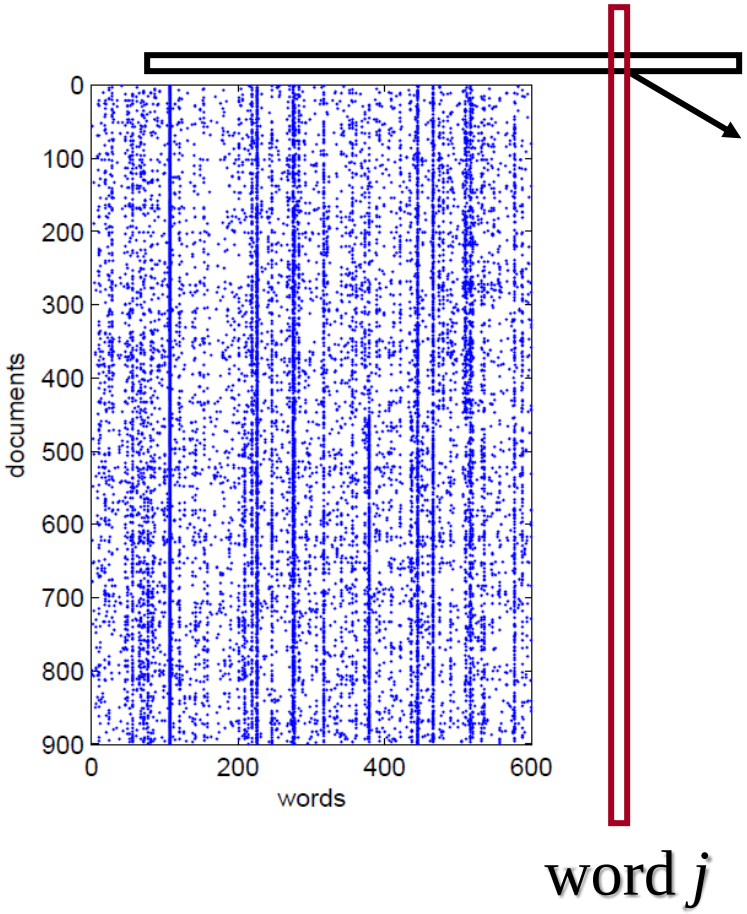
The board also detailed export registrations for sub-products, as follows....

→

word	frequency
grain(s)	3
oilseed(s)	2
total	3
wheat	1
maize	1
soybean	1
tonnes	1
...	...

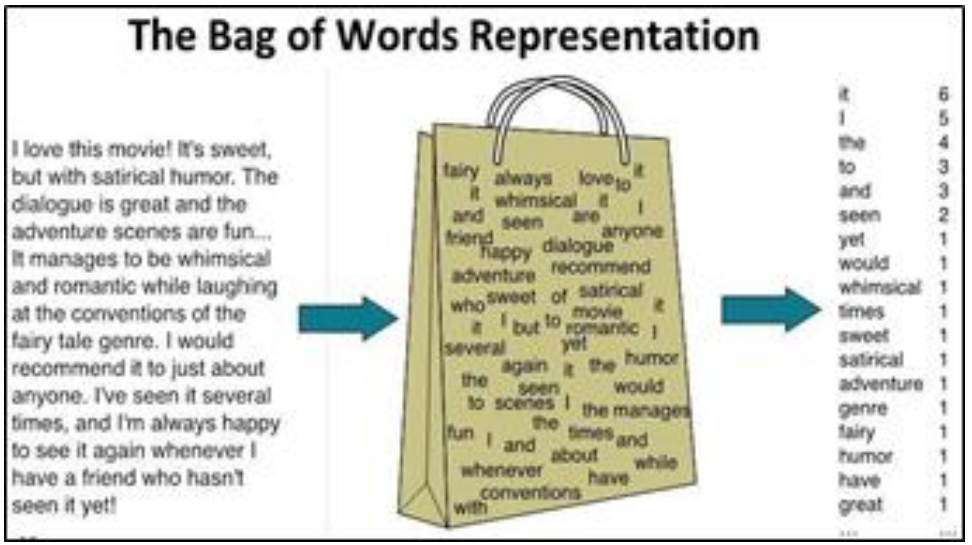
Bag of word representation: Represent a text as a vector of word *frequencies*.

‘Bag of words’ representation of text



Frequency (i,j) = j
in document *i*

Document-term matrix



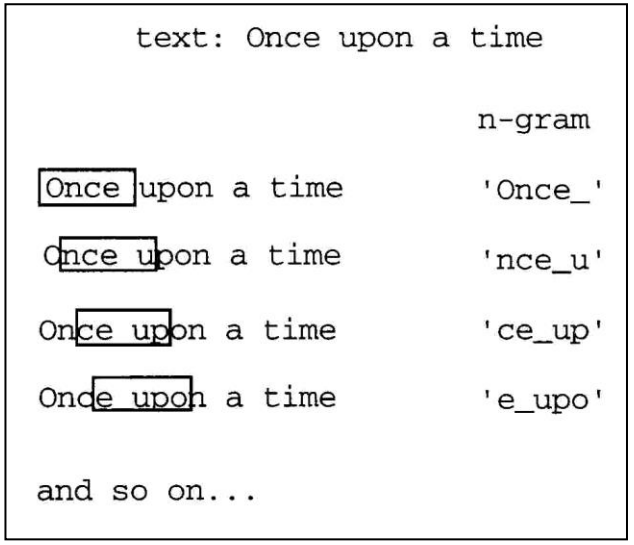
‘Bag of words’ representation of text

Variants include:

- How to weight a word within a document (boolean, tf*idf, etc.)
  - Boolean: 1 is the word i is in doc j, 0 else
  - Tf*idf and others: the weight is a function of the word **frequency** in the document, and of the frequency of documents whith that word
- What is a “word”:
  - single, inflected word (“going”),
  - lemmatised word (going, go, gone→go),
  - N-Grams.

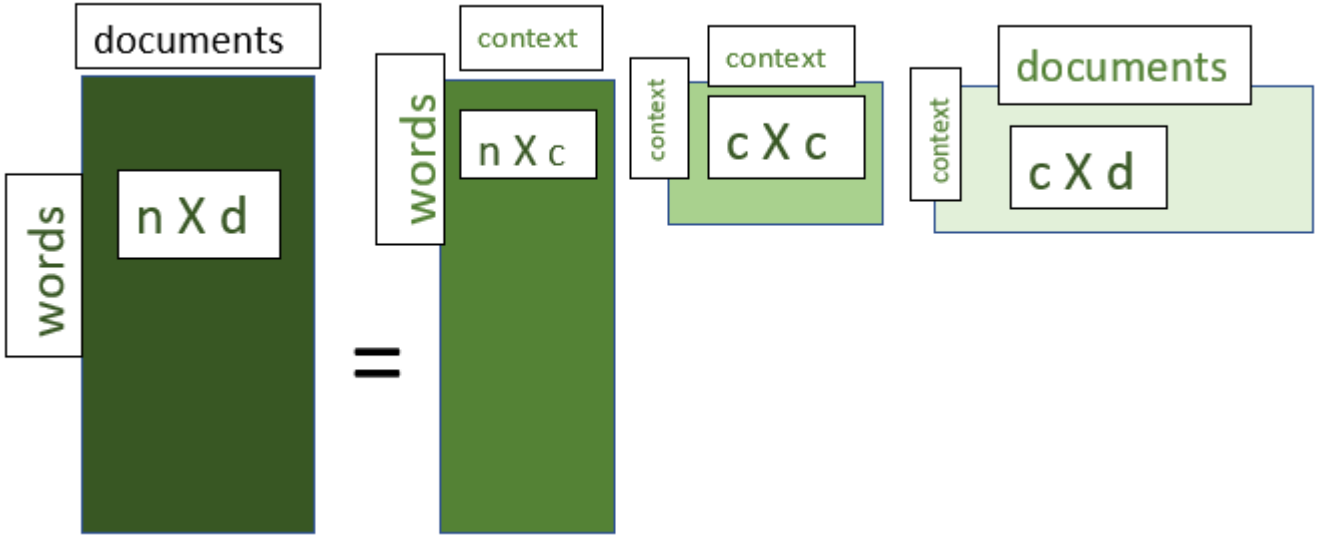
N-Grams

- *N*-gram is a connecting sequence of *N* symbols from a text
- In this approach a text is considered as a stream of characters and then it is broken down to substrings of length *N*
- It is remarkably resistant to textual errors, and no linguistic knowledge is needed.



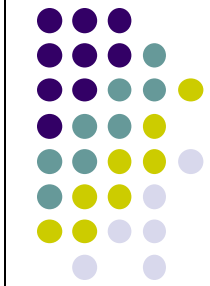
Embedding: Latent Semantic Analysis

Singular Value Decomposition of the cooccurrence matrix



- Compactional cost scales quadrigal for n x m matrix
- Bad for millions of words or documents
- Hard to incorporate new words or documents

Limitations of Bag-of-Words



The model is straightforward and proposes much flexibility in the adaptation to data. The methodology has been successful in predicting problems like language modeling and texts classification.

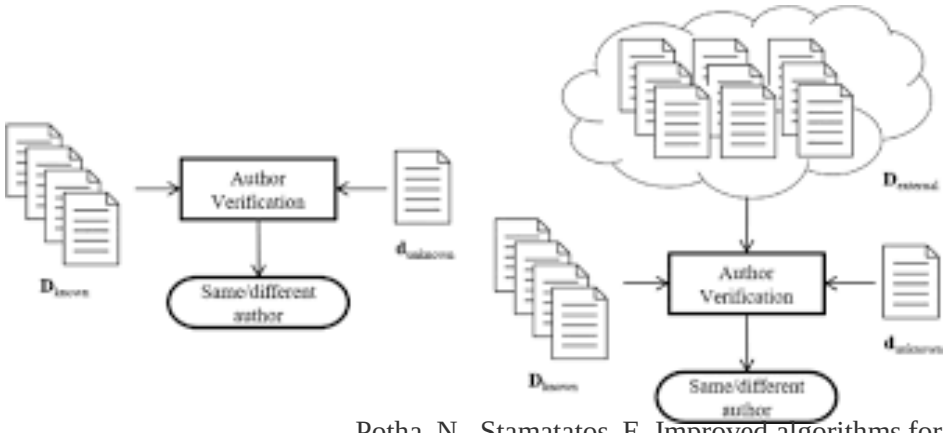
Nevertheless, there are several shortcomings, such as:

- **Vocabulary**: The vocabulary has to be carefully designed to manage the size, which impacts the sparsity of the document representations.
- **Sparsity**: Each dimension in this sparse representation corresponds to a "term" lying in the vocabulary.
- **Semantic**: Bag of word models does not respect the semantics of the words.
- **Meaning**: Discarding word order ignores the context and meaning of words in the document (semantics). Context and meaning can offer a lot to the model.

Author Verification

Verification models differ with respect to their view of the task.

- The **intrinsic** approach operates only with the provided texts (of acknowledged authorship and the questioned one) and leads to a one-class classification problem. Such methods are usually robust and fast while they do not require any external resources.
- **Extrinsic** verification models try to transform the verification task into two class classification task by bearing in mind external documents representing the negative class. They are usually found to be more effective than intrinsic methods



Potha, N., Stamatatos, E. Improved algorithms for extrinsic author verification. *Knowl Inf Syst* **62,** 1903–1921 (2020).

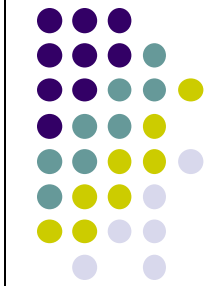
Imposters Method

	Impostor	X	Y
4-gramm	Frequency	Frequency	Frequency
bad!	F11	F12	F13
bank	F21	F22	F23
good	F31	F32	F33
text	F41	F42	F43
the_	F51	F52	F53
view	F61	F62	F63

The method checks whether X is more similar to Y than each of the imposters taken from a given collection, such that the evaluation is performed resting upon a collection of randomly chosen feature subsets (for instance, marked in red). The "similarity" of X to Y is the fraction of the scores where X and Y are "closer" to each another than to the imposter set.

(**Koppel, M. and Winter, Y.** (2014). **Determining if two documents are written by the same author.** ***Journal of the Association for Information Science and Technology*****, 65(1): 178–87)**

Similarity Measures



A similarity measure is a function that computes the *degree of similarity* between two vectors.

- Similarity between vectors for the documents is computed as the vector inner product (a.k.a. dot product)
- Chi-Squared distance
- Canberra distance
- Jensen–Shannon divergence
- Second order distance

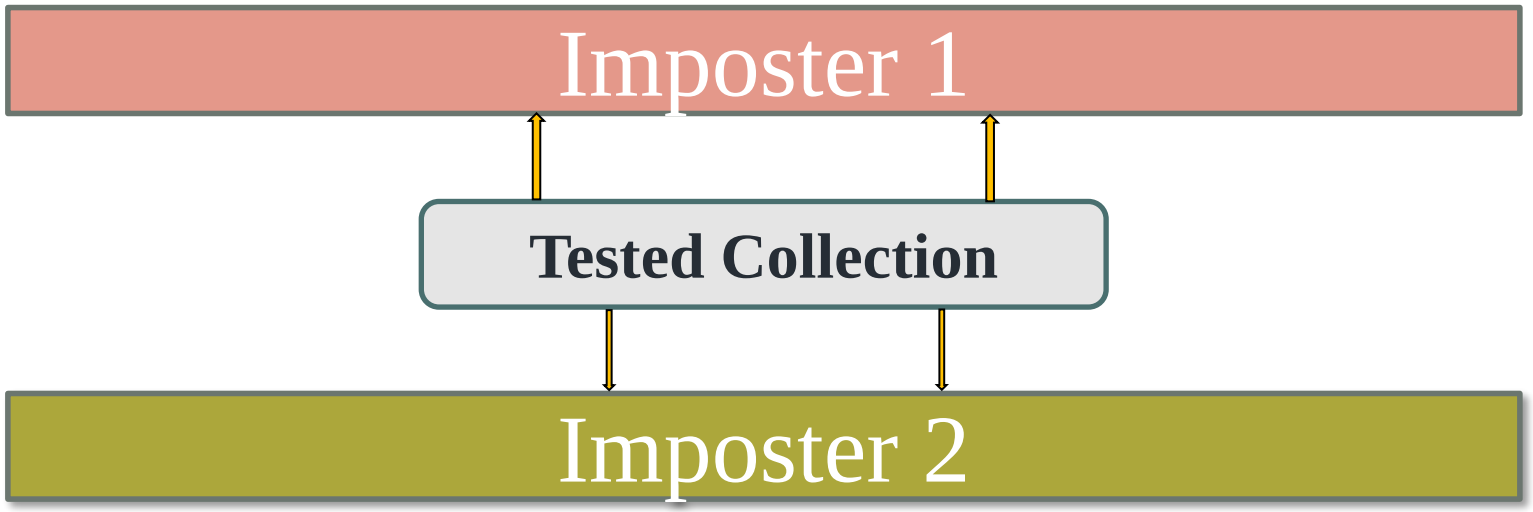
Method Weaknesses

The method inherits all the listed disadvantages of the mentioned earlier representation:

- The method does not respect the semantics of the N-grams ignoring their order. Discarding the order ignores the context and meaning in the document (semantics).
- The method compares merely one-dimensional marginal distributions of N-grams.
- N-grams distributions can be estimated appropriately just for sufficiently long texts. Thus, small patterns are not captured.

Deep “imposters” projections method

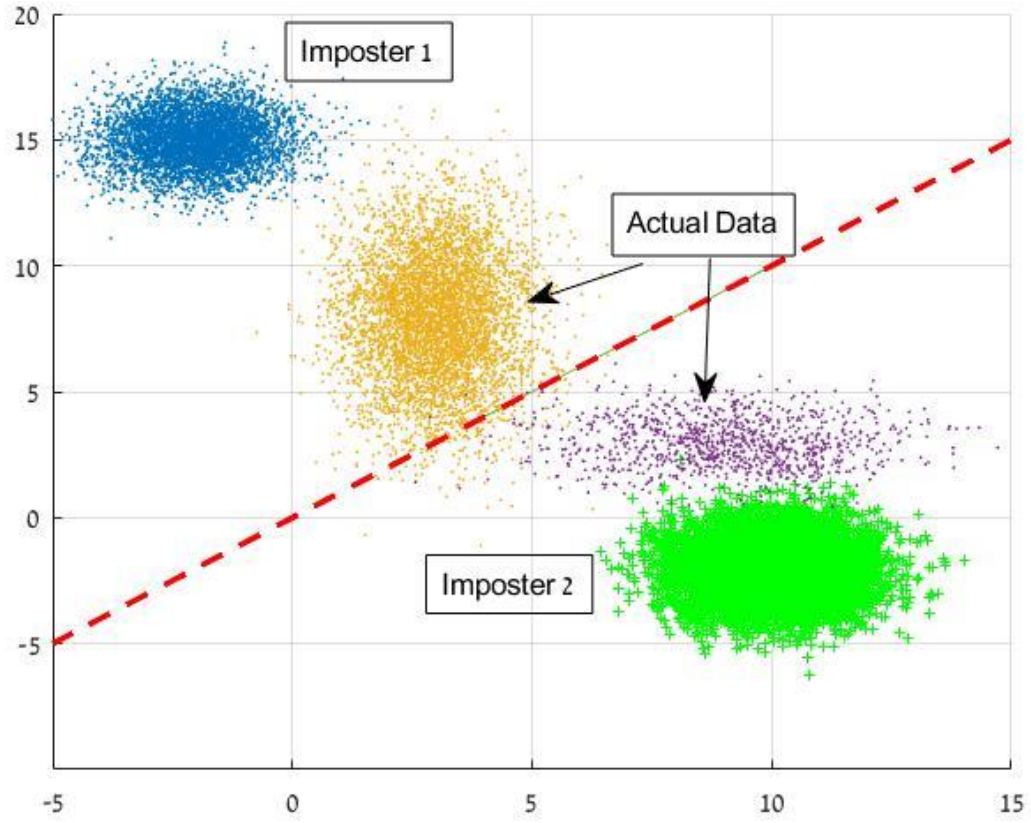
The method is based on labeling of sequential small fractions of the texts in the tested collection through a deep network trained on an imposters' pair.



The method is inspired by methodology proposed in:

Z. Volkovich, A Short-Patterning of the Texts Attributed to Al Ghazali: A “Twitter Look” at the Problem, Mathematics 2020, 8(11), 1937

Method Exemplification

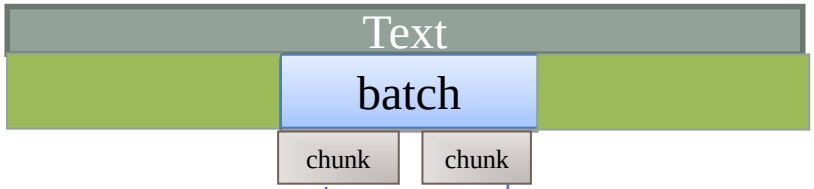


- At each step, a deep network is trained and provides a separating manifold (a red dotted line).
- A collection under consideration is classified according to the splitting rule generated by the network and transformed into a signal.

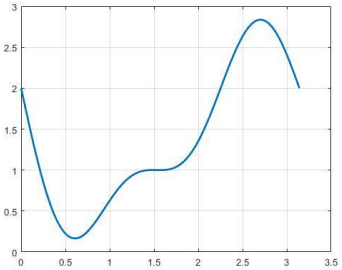
Deep “imposters” projections method

- The imposters' and the tested collections are divided in small bathes such that each one of them is divided in turn into small chunks like "tweets".
- Each chunk is assigned by the trained network to one of the imposters (0 or 1).
- The whole batch receives a score being the average of its chunks scores producing a signal representation.

..

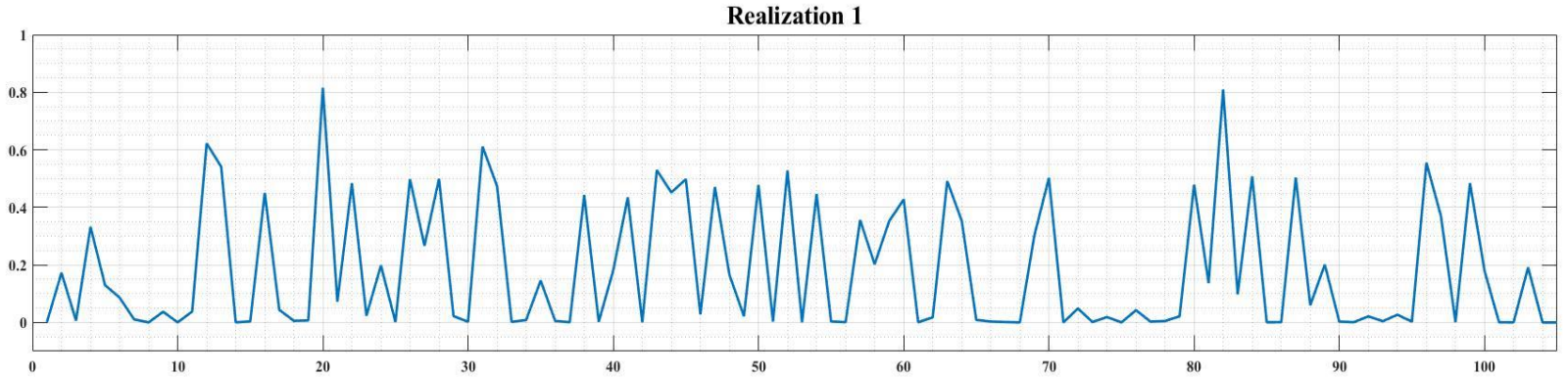


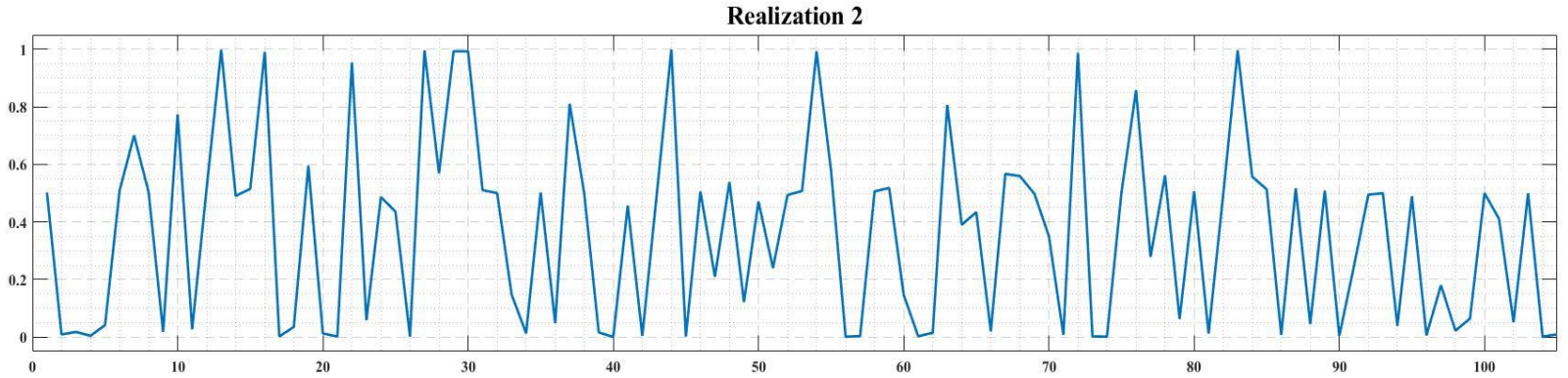
Mean score of a batch.



A signal representation

Examples of signals obtained for “THE TRAGEDY OF HAMLET PRINCE OF DENMARK”

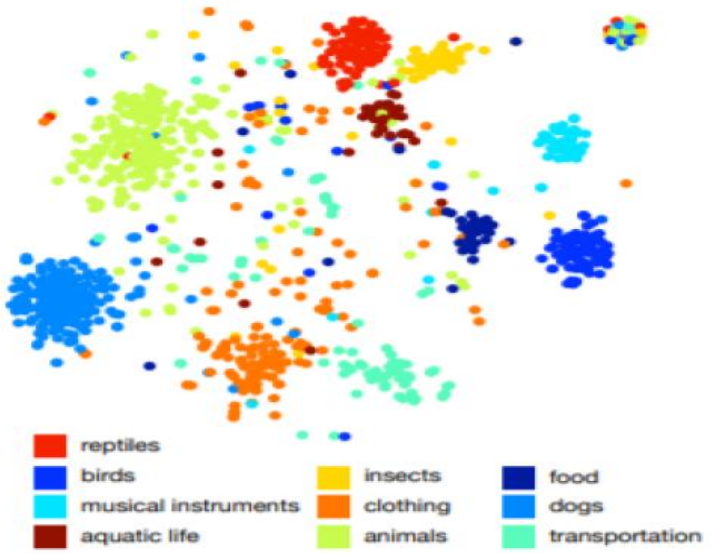




Representations are calculated for two different imposters pairs.

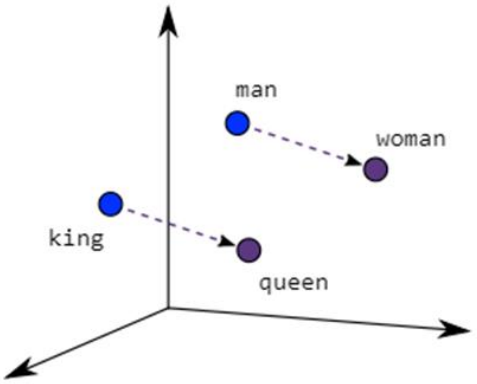
Word Embeddings

- Word embedding is a suite of language modeling methods presenting words of a given glossary in digital vector spaces, usually having high dimensionality.
- This incredibly valuable technique preserves the essential semantic and syntactic information.
- Popular off-the-shelf word embedding models in use today are:
  - Word2Vec (by Google)
  - GloVe(by Stanford)
  - FastText (by Facebook)
  - ELMo(AllenNLP's)
  - Transformers



Word2vec

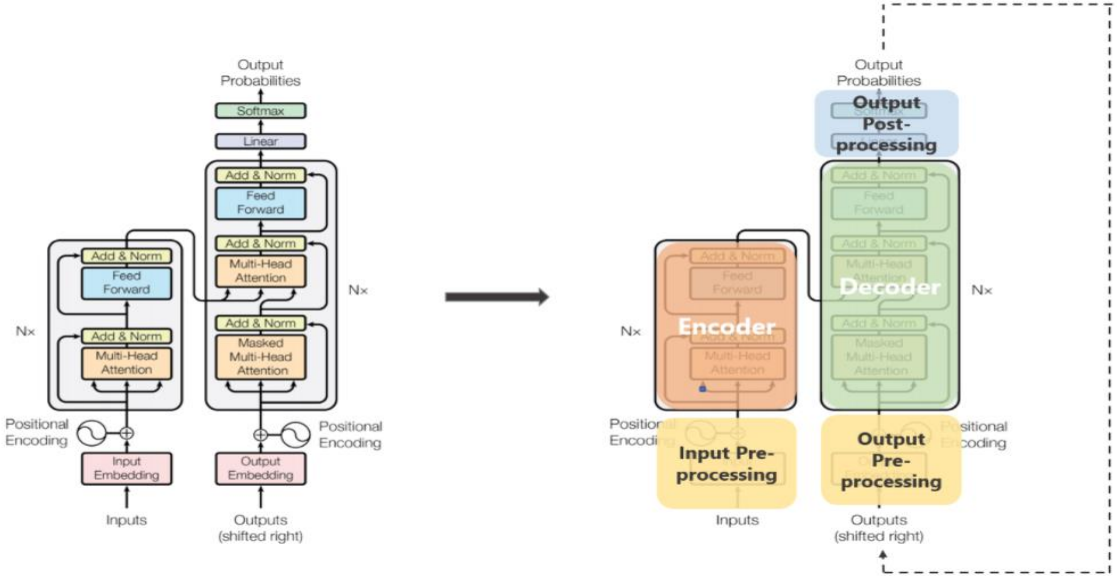
- Key idea: The word vector can predict surrounding words.
- Word2vec: as originally described (Mikolov et al 2013), a NN model using a two-layer network (i.e., not deep!) to perform dimensionality reduction.
- Faster and can easily incorporate a new sentence/document or add a word to the vocabulary.
- Very computationally efficient, good all-round model (good hyper-parameters already selected).



$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“aunt”}) - W(\text{“uncle”})$$

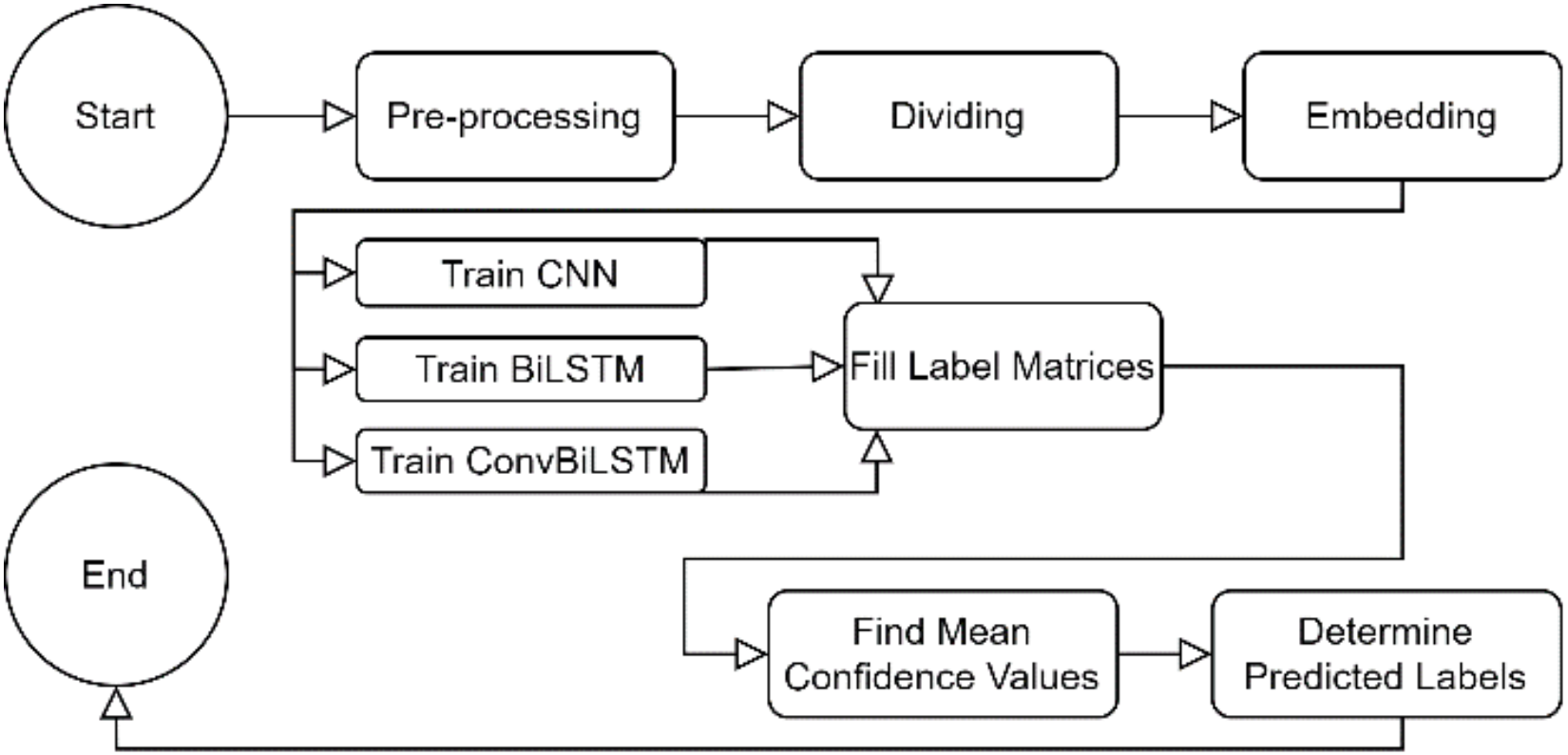
$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“queen”}) - W(\text{“king”})$$

Transformers

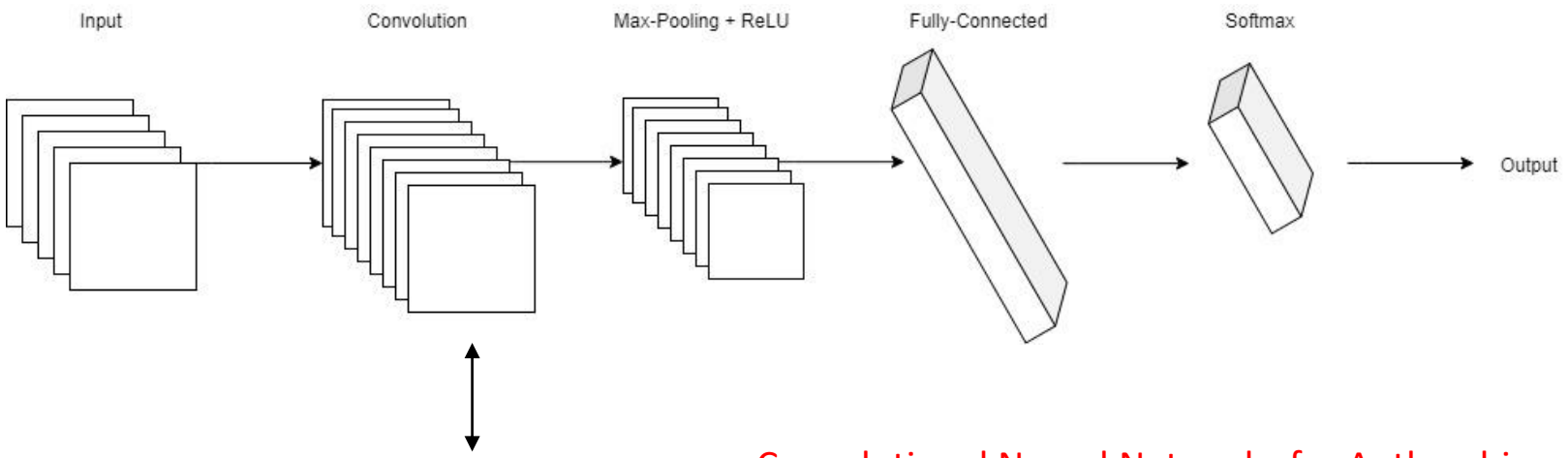


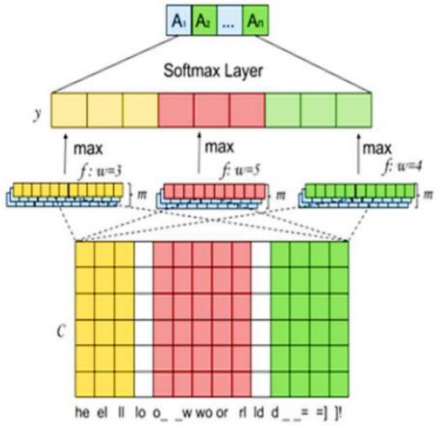
- The encoder transforms an input sequence into a sequence of continuous representations input into a decoder.
- The decoder obtains the output of the encoder and the decoder output at the previous step to produce a resulted sequence.
- This is a currently developed technique providing efficient word embeddings.

General Model Flow Chart



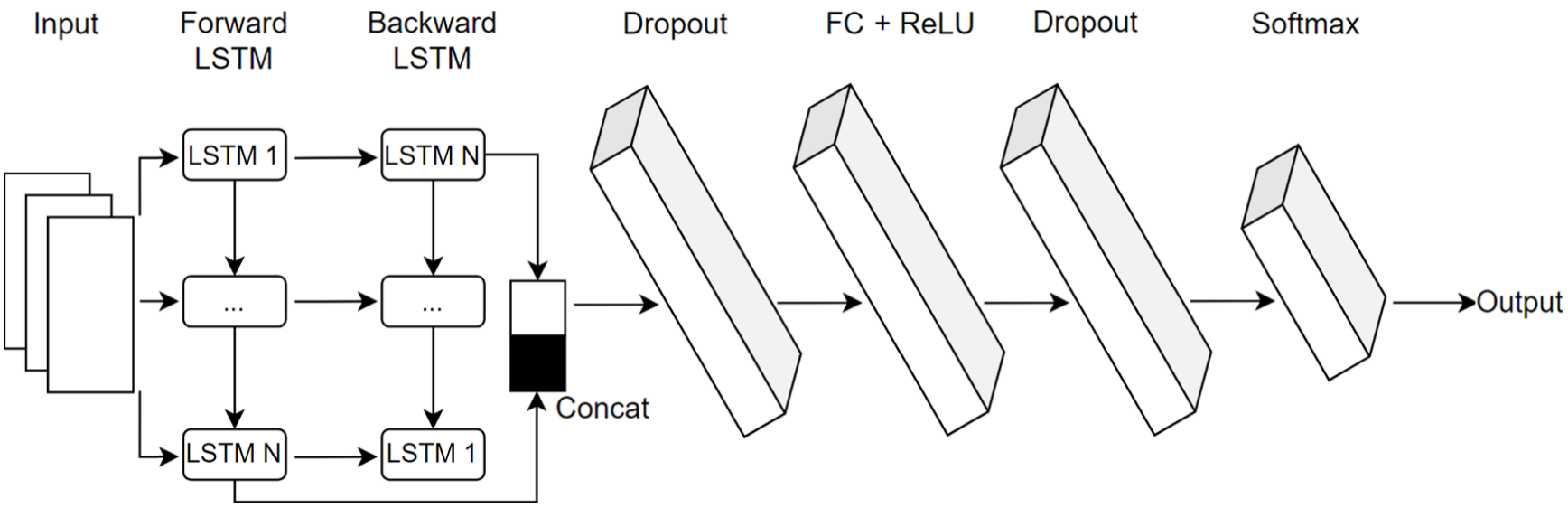
CNN Architecture



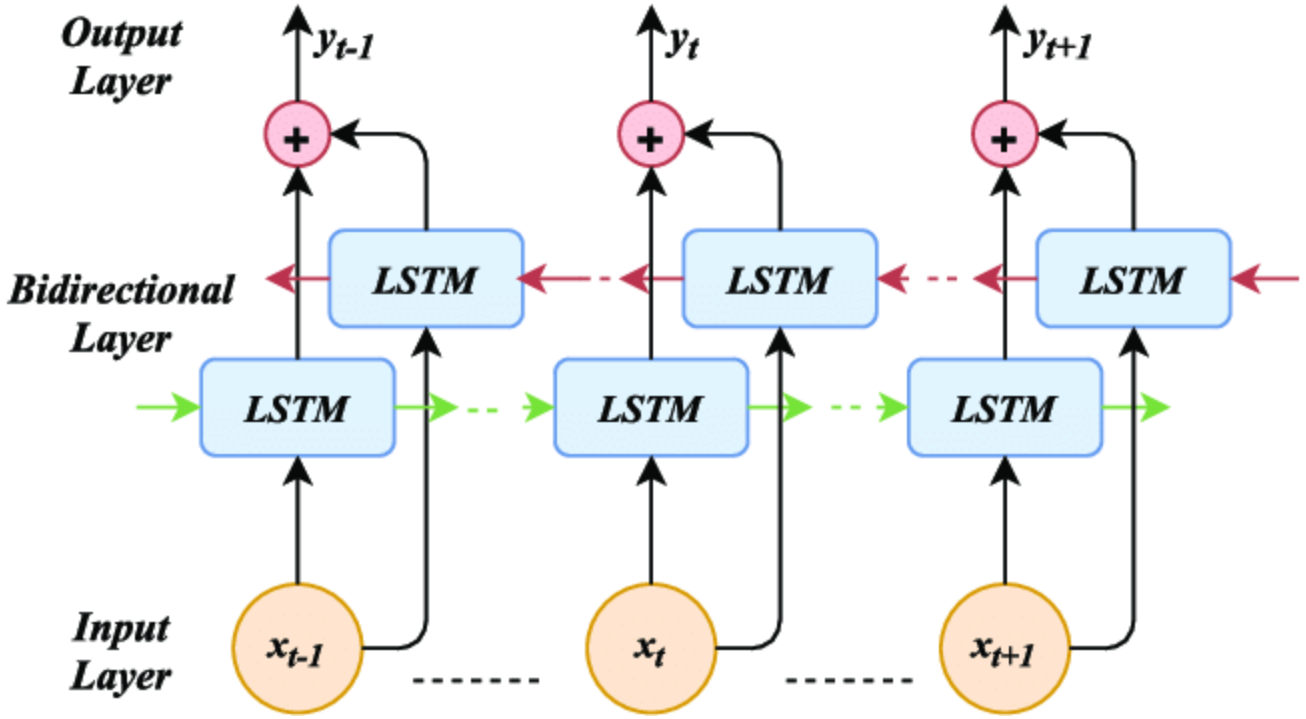


Convolutional Neural Networks for Authorship Attribution of Short Texts, Prasha Shrestha, Sebastian SierraFabio,
A. González, Thamar Solorio.
Conference: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers

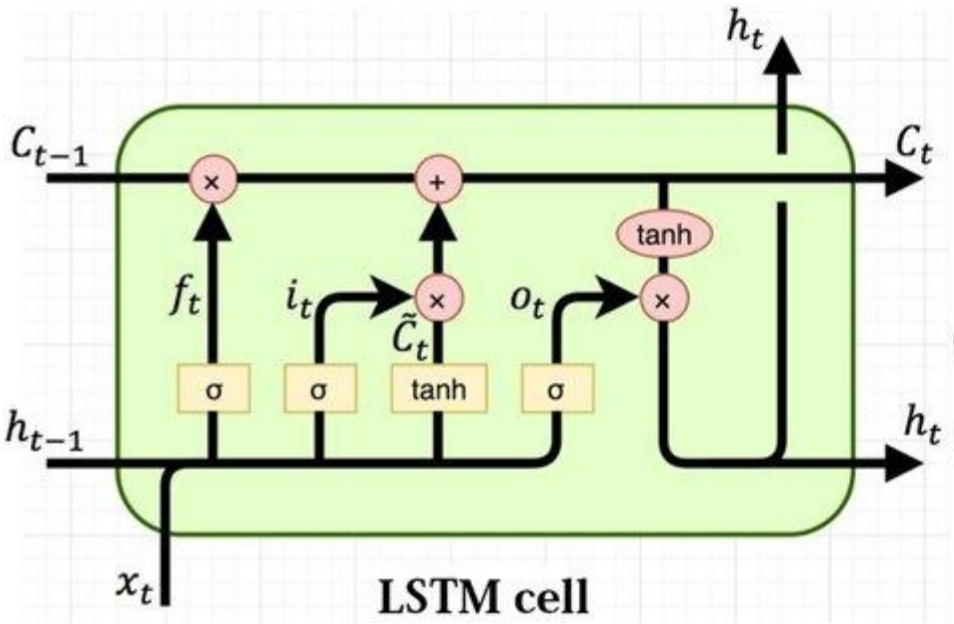
BiLSTM Architecture



Bidirectional LSTM



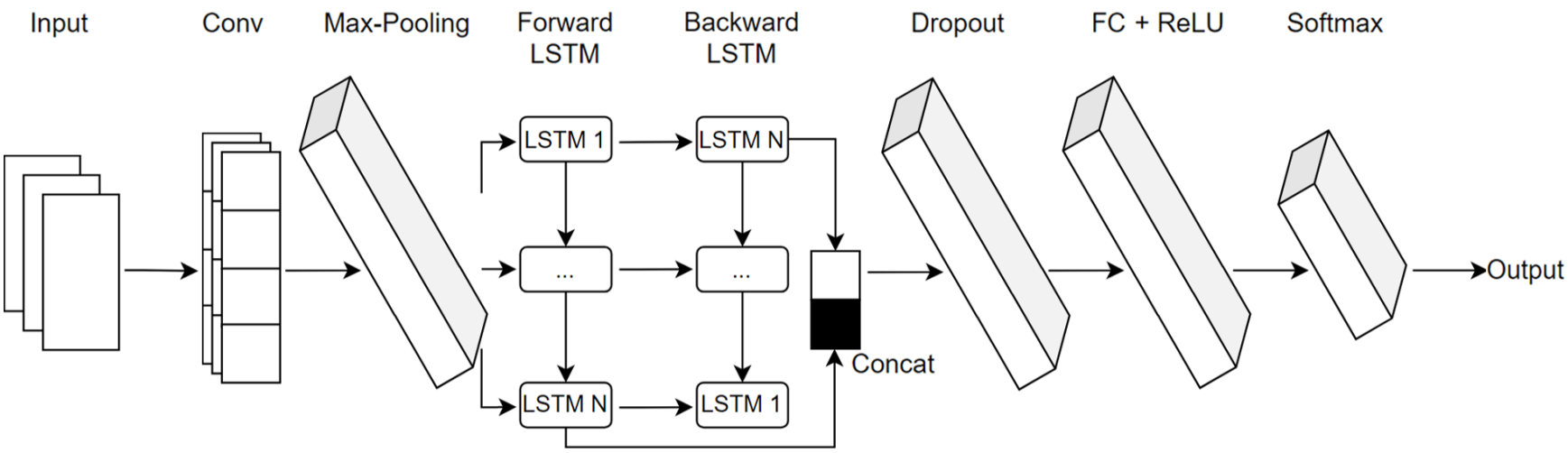
LSTM cell



$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right)$$
$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right)$$
$$o_t = \sigma\left(x_t U^o + h_{t-1} W^o\right)$$
$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right)$$
$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$
$$h_t = \tanh(C_t) * o_t$$

LSTM networks are suited to classifying and making predictions based on dynamic data.

Conv-BiLSTM Architecture



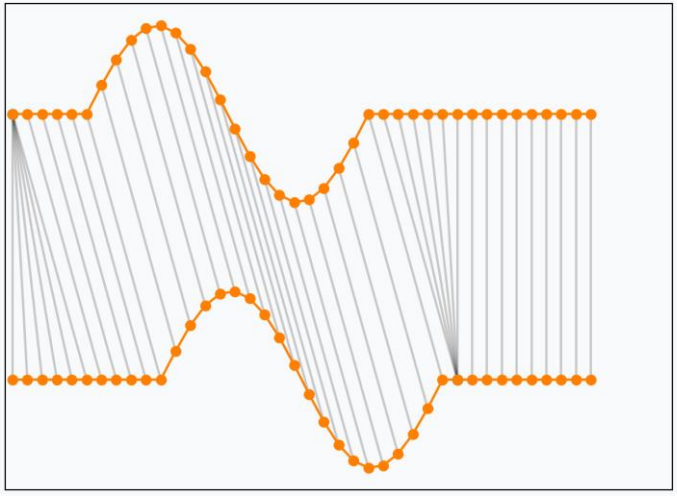
Like “Shakespeare Apocrypha”

The Shakespeare Apocrypha is a collection of creations sometimes questionable or not attributed to William Shakespeare.

- The tested collection contents 52 creations attributed to William Shakespeare downloaded from Project Gutenberg authority https://www.gutenberg.org/.
- Imposters collections of 11 authors are downloaded from the same authority:
  - Arthur C Clarke
  - Benjamin Jonson
  - Charles Dickens
  - Christopher Marlowe
  - Francis Bacon
  - Geoffrey Chaucer
  - Harry Potter
  - Jane Austen
  - John Galsworthy
  - Isaac Asimov
  - Robert Sheckley

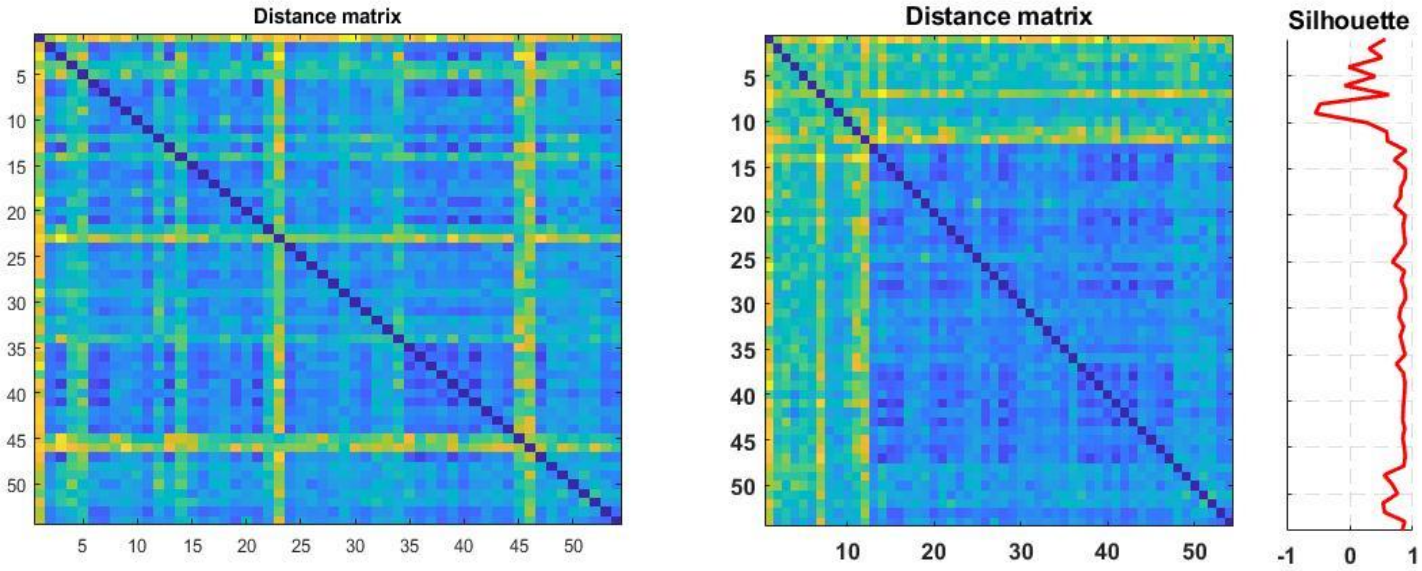
Procedure

1. Upload (The Test Collection)
2. Upload (The Imposters)
3. For each Imposters pair do:
   - Pre-Processing
   - Train a ConvBiLSTM Net
   - Assign chunks of the Test Collection texts and obtain signal representation
   - Calculation of a matrix of pairwise DTW distances.



Dynamic Time Warping is equivalent to minimizing the Euclidean distance between the aligned time series under all admissible temporal alignments.

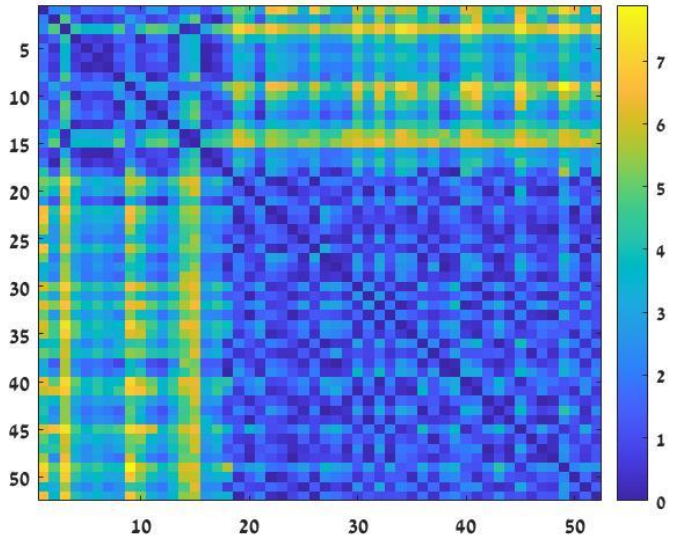
Pairwise DTW distances



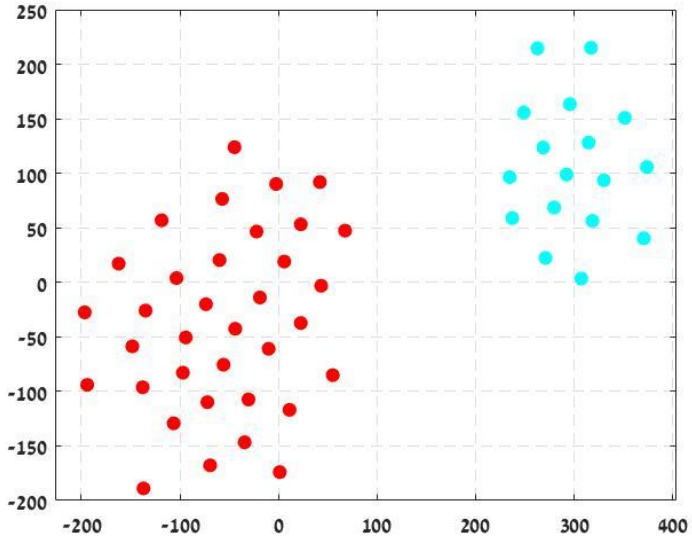
An example of DTW distances matrix(the left panel)and its clustered and permuted into 2 clusters version (the right panel). A core corresponding to the genuine creations manifests.

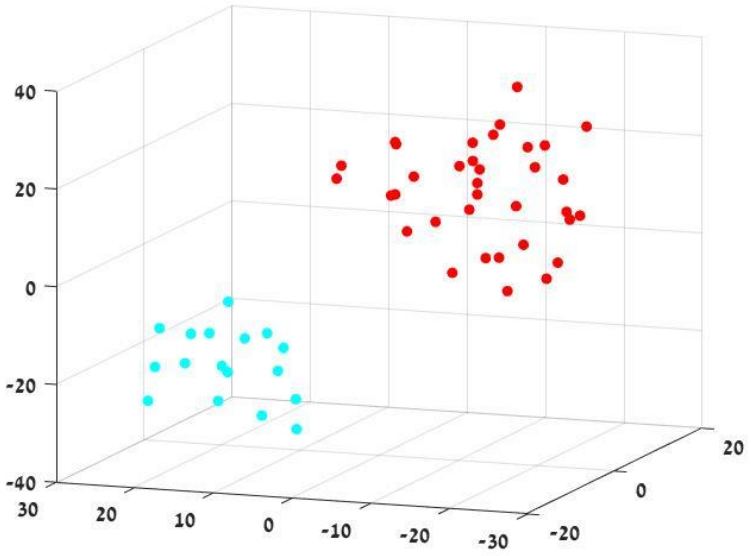
Procedure Summarizing

4. Summarizing
   - Collect Random forests probability in 1000 iterations of the DTW matrixes
   - Collect DTW matrix clustering
5. Grouping the results into clusters



T-SNE projections of the results





There is an evident partition of all documents in two clusters. The red category consists supposing of the genuine William Shakespeare creations.

16 Suspected Pseudo

Name
'Arden of Feversham'
'The Puritaine Widdow by Shakespeare'
'Locrine Mucedorus by Shakespeare'
'A Yorkshire Tragedy by Shakespeare'
'The Merry Devill of Edmonton by Shakespeare'
'Sir John Oldcastle by Shakespeare'
'The Two Noble Kinsmen by Shakespeare'
'Sir Thomas More by Shakespeare'
'THE MERRY WIVES OF WINDSOR'
'A MIDSUMMER NIGHT_S DREAM'
'KING HENRY VI part I'
'KING HENRY V'
'Comedy of Errors'
'KING JOHN'
'King Henry IV, the First Part'
'THE TRAGEDY OF CORIOLANUS'

Abel Lefranc suggests that the "Merry Wives of Windsor" is based on events in Derby's lifespan

Abel Lefranc suggests that A Midsummer Night's Dream was written for Derby's own marriage to Elizabeth de Vere

Neal Fox, Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate

Heterogeneous Authorship in Early Shakespeare and the Problem of Henry V. THOMAS MERRIA

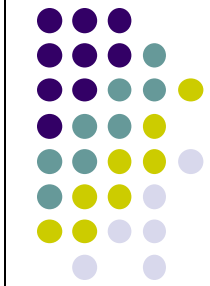
We, please leave the results to Shakespeare experts to clarify the causes such a partition.

Abu Hamid Al Ghazali

- The outstanding Islamic jurist, theologian, and mystical thinker Abu Hamid Al Ghazali (1058–1111) is one of the most significant Muslim Sufis, whose ideas are prominent and convincing not only in the Muslim world.
- According to the Hadith predicting the arrival of Islam's renewer once every century, the Arab community perceived Al Ghazali as the renewer of Islam's fifth century. The Shafi'i jurist al-Subki claimed,"If there had been a prophet after Muhammad, Al-Ghazali would have been the man".
- His deep creativity and popularity have caused many counterfeits and imitations.

Z. Volkovich, A Short-Patterning of the Texts Attributed to Al Ghazali: A "Twitter Look" at the Problem, Mathematics 2020, 8(11), 1937

Material Al Ghazali

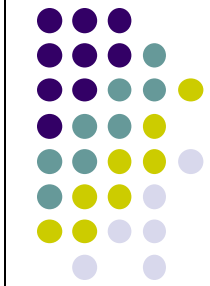


- The source collection (Cl_0) (the First "Imposters" set) contains text from Al Ghazali's most significant work, *Iḥyāʾ ʿulūm al-dīn(***The Revival of the Religious Knowledge**), downloaded from the site http://ghazali.org/ihya/ihya.htm as a collection of 41 files with a total size of 8.5 MB. It is regarded as one of his chief works and a classic introduction to the pious Muslim's way to God.

- The alternative collection (Cl_1)(the Second "Imposters" set) includes 9 texts definitely attributed as not written by Al Ghazali's, with a total size of 1.0 MB;

- The test collection is composed of
  - 7 texts agreed upon as written by Al Ghazali;
  - 2 texts agreed upon as not written by Al Ghazali (Pseudo-Ghazali);
  - A book with questionable authorship: *Mishakat al-Anwar* (*Niche of Lights*).

Balancing procedure



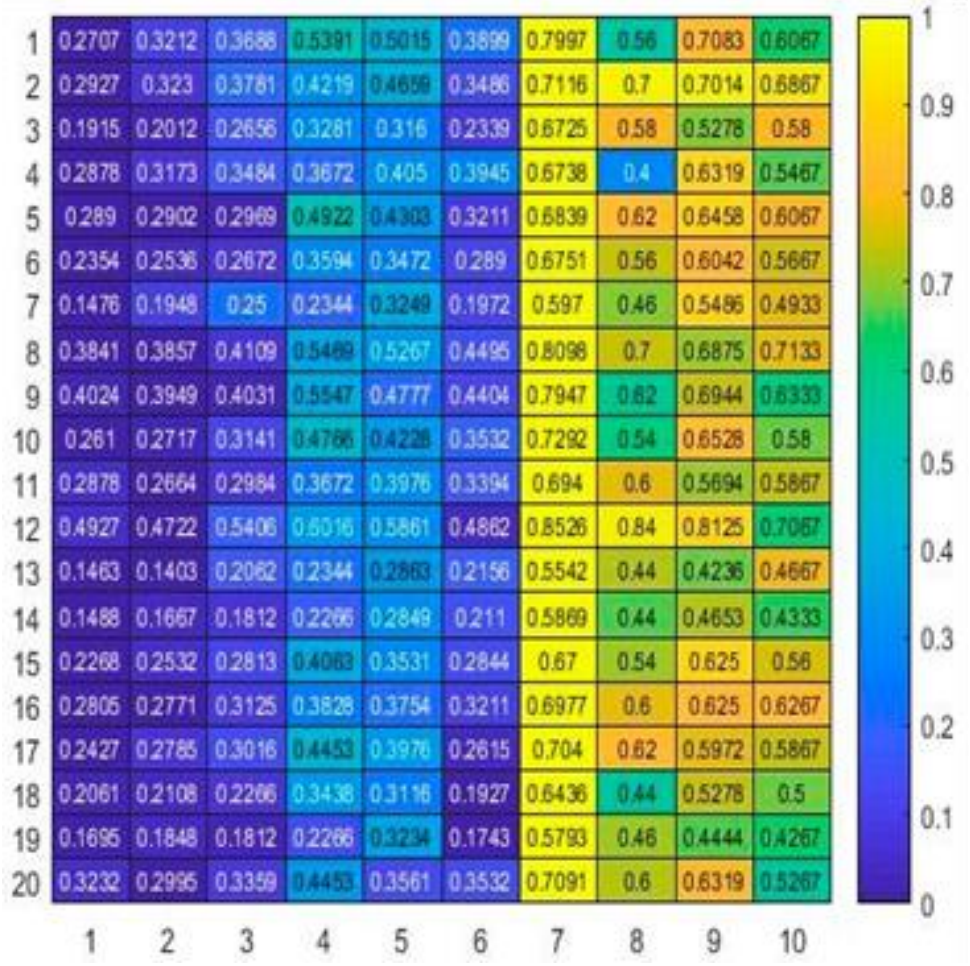
Due to significant difference in the sizes of Cl_0 and Cl_1 (approximately 8.5 and 1.0 MB), a balancing procedure is applied.

1. A sample S from Cl_0 with the undersampling rate $2^* |Cl_1|$ without replacement is taken from Cl_0
2. A sample V is obtained 3 times replication of Cl_1
3. S and V are replicated 3 times

The last step is augmentation used in building convolutional neural networks for increasing the size of the training set without acquiring new data.

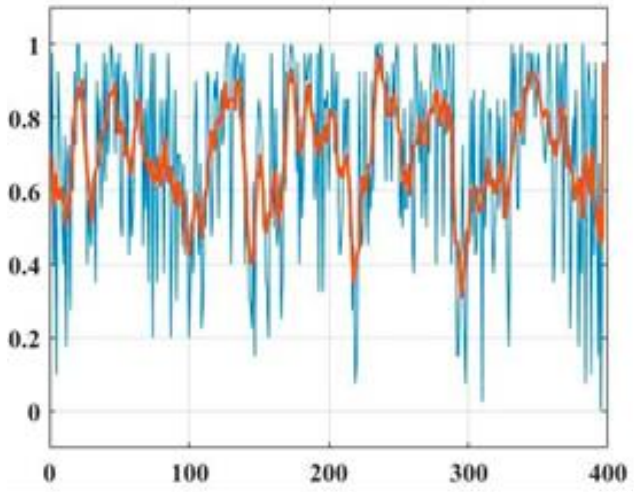
The resulted collections S_0 and S_1 are an input to a network.

A heat map of the mean values attained in 20 iterations

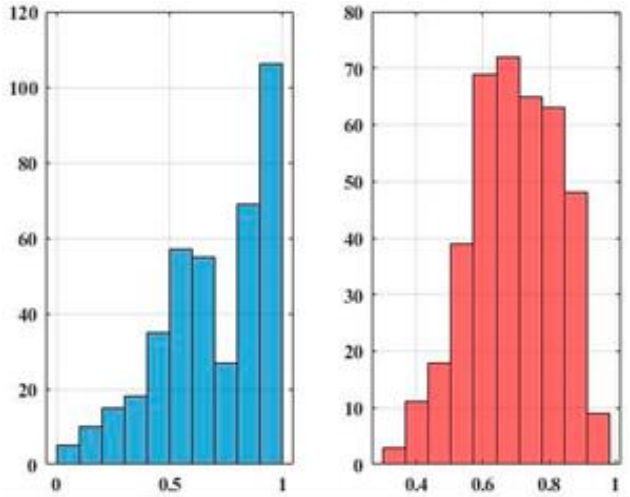


The Parula color map in the “scaled rows” fashion

“Tahafut al-Falasifa”(*The Incoherence of the Philosophers)(a landmark 11th-century work)*



A digitally averaged representation of “*Tahafut al-Falasifa”* batches over 20 iterations



Histograms of the chunked scores of “Tahafut al-Falasifa” for the averaged profile (the left panel) and its smoothed version (the right panel).

Reminder: 0 means the inherent Al Ghazali style of a batch , and 1 - NO

- Both distributions have a negative skew, specifying a long left tail, the left asymmetry of a distribution around its mean. About 20% of the data are smaller than 0.5 (**inherent Al Ghazali)** in the left panel and about 8% (**inherent Al Ghazali)** in the red one.
- Thus, it is possible to conclude that the vast part of the considered manuscript *Tahafut al-Falasifa,* is not written in the inherent Al Ghazali style.
- The acceptable historical version is that it is a common creation with a student of the Asharite school of Islamic theology.
- Maybe, the student was transcribing Al Ghazali’s ideas in his own words

Hebrew Bible

- A traditional base belief in Judaism is the Mosaic authorship of the Torah, being the first five books of the Hebrew Bible dictated to Mosaic, according to the main Jewish acceptance, by God himself on Mount Sinai.
- The Torah collection includes
  - Bəreshit (,בְּרֵאשִׁית literally "In the beginning")—Genesis, from Γένεσις (Génesis, "Creation")
  - Shəmot (,שְׁמוֹת literally "Names")—Exodus, from Ἔξοδος (Éxodos, "Exit")
  - Vayikra (,וַיִּקְרָאliterally "And He called")—Leviticus, from Λευιτικόν (Leuitikón, "Relating to the Levites")
  - Bəmidbar (,בְּמִדְבַּרliterally "In the desert [of]")—Numbers, from Ἀριθμοί (Arithmoí, "Numbers")
  - Dəvarim (,דְּבָרִיםliterally "Things" or "Words")—Deuteronomy, from Δευτερονόμιον (Deuteronómion, "Second-Law")
- In that sense, Torah means the same as Pentateuch or the Five Books of Moses.
- Questions regarding the Torah authorship continuously arise. E.g., whether Moses could have written the account of his death in Deut. 34:5-12.
- The medieval Jewish commentators such as Abraham Ibn Ezra announcement other inconsistencies as well.

Imposters collections

I. (Cl_0) **Source set of 3 Pentateuchal books traditionally agreed to be written by *Moses*:**

ENGLISH	HEBREW	WORD COUNT
Genesis	(בְּרֵאשִׁית)	32K
Exodus	(שְׁמוֹת)	26K
Numbers	(בְּמִדְבַּר)	25K

I. (Cl_1) **Impostor set of 3 books traditionally agreed not to be written by *Moses*:**

ENGLISH	HEBREW	WORD COUNT
Psalms	(תְּהִלִּים)	30K
Jeremiah	(יִרְמְיָה)	33K
Isaiah	(יְשַׁעְיָהוּ)	25K

Results

Three networks overall are used:

1.CNN

2.Bi-LSTM

3.ConvBi-LSTM

with Embedded through ELMo,GPT-2,Bert

Network	(1)		(2)		(3)	
Book						
Deuteronomy	**Not Moses**	**70.9%**	**Not Moses**	**73.1%**	**Not Moses**	**63.4%**
Job	**Not Moses**	**96.8%**	**Not Moses**	**98.2%**	**Not Moses**	**81.7%**
Leviticus	**Moses**	**99.8%**	**Moses**	**99.1%**	**Moses**	**80.5%**
Ezra	**Not Moses**	**84.1%**	**Not Moses**	**88.9%**	**Not Moses**	**78.8%**
Nehemiah	**Not Moses**	**75.5%**	**Not Moses**	**84.6%**	**Not Moses**	**66.9%**
Joshua	**Not Moses**	**66.7%**	**Not Moses**	**70.6%**	**Not Moses**	**74.0%**

Comparison

Document	*Initial Attribution*	*Predicted Attribution*
Deuteronomy	Moses (Controversial)	73.4 Not Moses
Job	Moses (Controversial)	97.7 Not Moses
Leviticus	Moses	03.5 Moses
Ezra	Not Moses	84.3 Not Moses
Nehemiah	Not Moses	83.5 Not Moses
Joshua	Not Moses	75.3 Not Moses

- All documents that are initially not prescribed to Moses are correctly determined to be not written by Moses. Leviticus, a book widely agreed upon to have been written by Moses, is determined as written by Moses. However, both Deut. and Job, which are considered somewhat controversial among scholars, are determined to be not written by Moses. Attributions were equal across all 3 proposed deep neural networks.
- The obtained results demonstrate that Leviticus was undoubtedly written by the same author of the other three widely accepted Torah books – Moses, while Deut. and Job were probably not.

Questions?