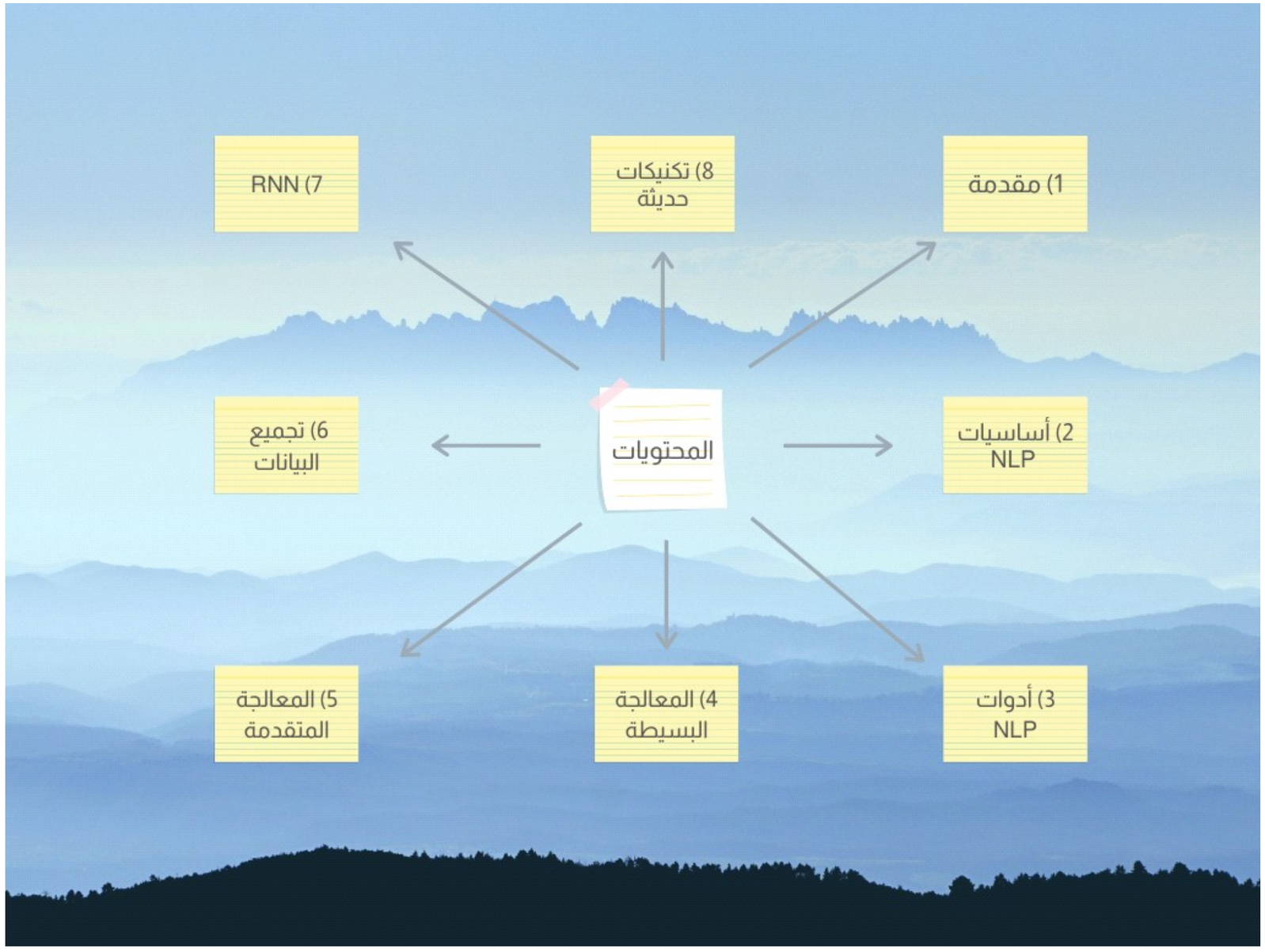
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

القسم السابع : الشبكات العصبية المتكررة

الجزء الرابع : Rec NN\TNN

==========

الشبكات العصبية المعادة Recursive Neural Network

و هي نوع من الشبكات العصبية التي يتم استخدامها بشكل اساسي مع النصوص , والتي يكون لها شكل هرمي مختلف عن الشكل التقليدي , وأحيانا ما يطلق عليها الشبكات العصبية المشجرة Tree Neural Network و يرمز لها TNN

و قبل ان نتحدث عنها , علينا ان نتناول العيوب التي قامت هي بتلافيها , فمن المشاكل التي تقابل موديل bag of words هي ان الترتيب بالغ الأهمية فكلمتي dog toy , toy dog مختلفتان في المعني

- Order is lost
- “Toy dog” vs. “Dog toy”

كذلك جملتي :

I love math and hate physics
I love physics and hate math

كلاهما سيحصلان علي نفس القيم في جدول BOW بينما هما مختلفان في المعني

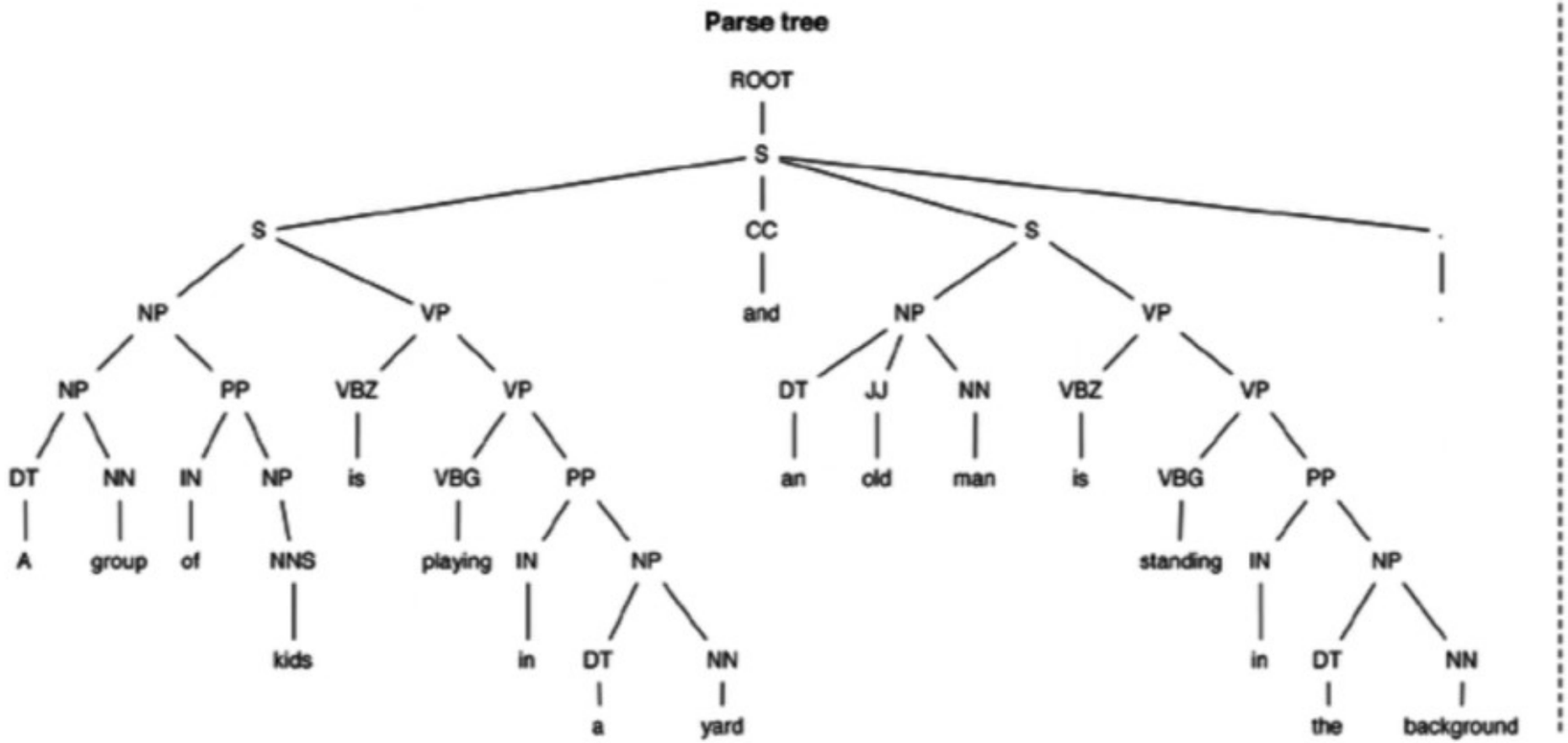
	I	Love	Math	And	Hate	Physics
Phrase 1	1	1	1	1	1	1
Phrase 2	1	1	1	1	1	1

كما أن RNN مشكلتها في الجمل الكبيرة , التي تعجز عن ربط الكلمة المطلوبة بعدد مخصص من الكلمات السابقة , فالإنسان بطبيعته لا يقرأ الجملة الكبيرة كلها مرة واحدة , و لكن يقسمها الي اجزاء صغيرة كي يفهمها

فلو كان لدينا جملة ما و هي :

A group of kids as playing in a yard and an old man is standing in the background .

فهذه الجملة يتم تقسيمها الي ما يشبه التقسيم الهرمي hierarchical , بحيث تتجمع الكلمات المتعلقة ببعضها البعض في حزم و أقسام , هكذا :



_

ففكرة ال TNN انها تقوم بعمل شبكة عصبية تتشابه في تصميمها مع هذا الشكل الهرمي , كي تتمكن من تدريب البيانات الهرمية عليها
و علينا أن نتعرف علي الأنواع الأساسية للكلمات , وهي :

- أولا الـ Verb , و يرمز لها V وهي : الأفعال
- ثانيا الـ Noun , و يرمز لها N وهي: الاسماء , الأماكن , الأشياء , الأشخاص
- ثالثا الـ Determiner , و يرمز لها D وهي: وسائل الربط , مثل حروف الجر و ادوات الملكية و ادوات التعريف the ,a , an , your
- رابعا الـ verb phrase , و يرمز لها VP وهي : الجملة الفعلية
- خامسا الـ noun phrase , و يرمز لها NP وهي : الجملة الإسمية

و مثال مبسط لها , جملة

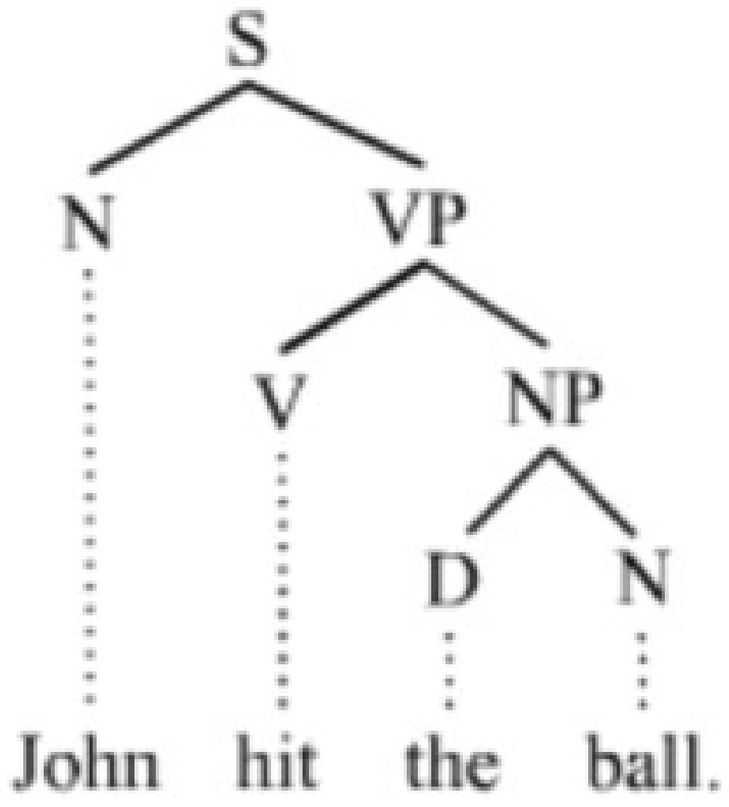
John hit the ball

فكلمة john هي N و hit هي V و the هي D و ball هي N

ثم نري أن هناك علاقة مباشرة بين the , ball فهما الاثنين يشكلان NP

ثم نري ان هناك علاقة بين فعل hit و جملة the ball و بالتالي يكونان معا VP

اخيرا نري أن john هو الفاعل أساس الجملة مع hit the ball لذا فهي ال sentence

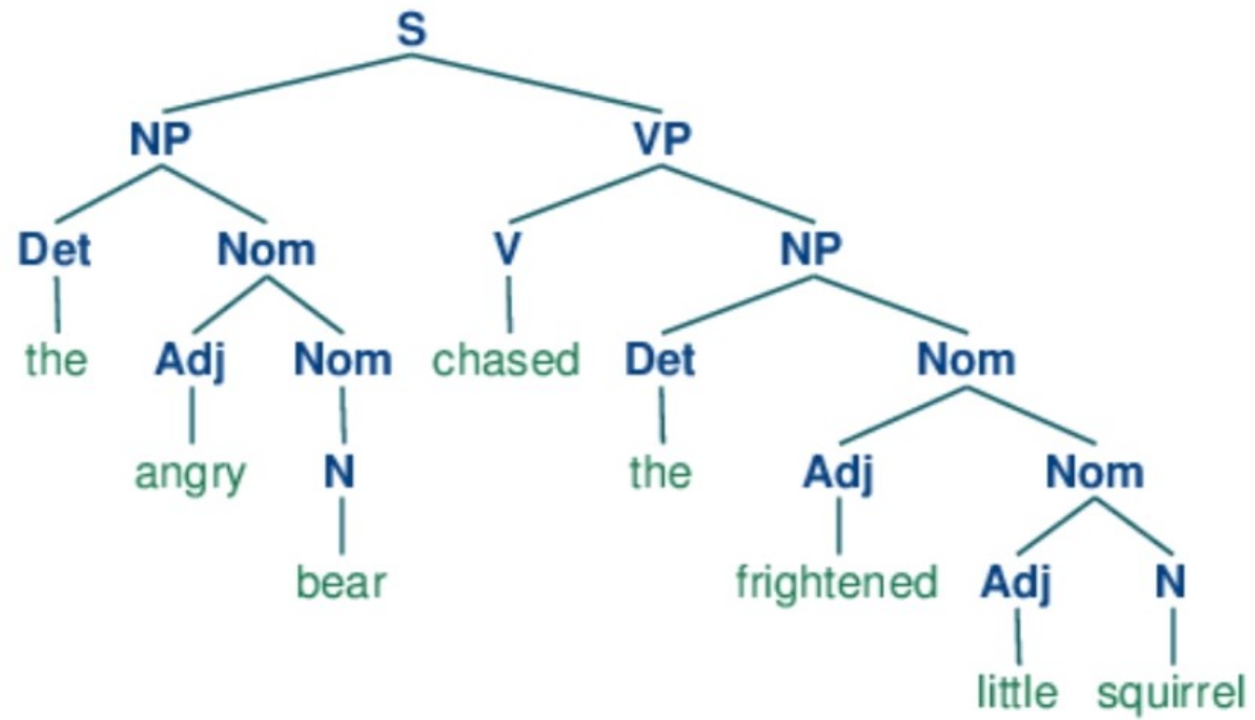


و اي ترتيب مختلف عن هذا لن يكون صحيحا , فلا توجد علاقة مباشرة بين hit the و حتي كلمتي john hit لا تشكلان معا معني كامل

و يكون الأمر أكبر مع جملة مثل :

The angry bear chased the frightened little squirrel

لتكون هكذا :



_

مع العلم ان طبيعة النصوص تجعل من الممكن عمل استبدال جزء كبير بكلمة واحدة , وان تظل محافظة علي المعني الأساسي , فيمكن استبدال الـ VP الحمراء هذه بكلمة أخري مثل

The angry bear chased the frightened little squirrel

The angry bear ran

او أن يتم استبدال الـ NP الخضراء بكلمة واحدة

The angry bear chased the frightened little squirrel

James chased the frightened little squirrel

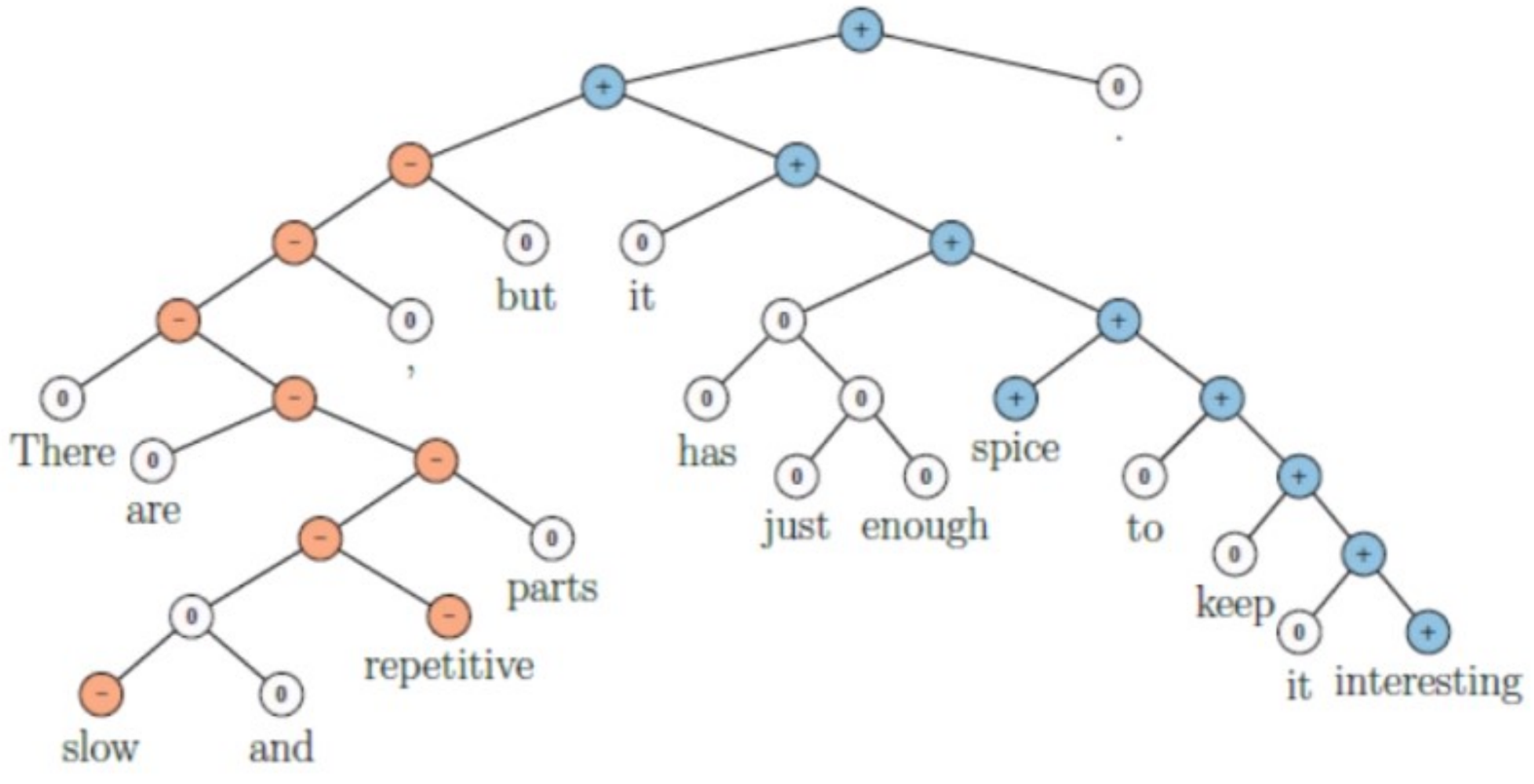
_

و هناك امثلة اكثر تعقيدا مثل استخلاص المعني من الجمل
فلو كان لدينا تعليقا علي فيلم , ونريد معرفة هل هذا التعليق سلبي ام ايجابي , فيمكن للشكل الهرمي ان يقوم باستخلاص معني كل كلمة و كل جملة علي حدة , لمعرفة المعني الإجمالي هل هو ايجابي ام سلبي

ففي الجملة التالية :

There are slow and repetitive parts , but it has just enough spice to keep it interesting .

نري أن الجزء الأول من الجملة بمعني سلبي, بينما الجزء الثاني معني ايجابي , فيمكن رسم شجرة هكذا :



و نري أن هناك كلمات و اجزاء ذات معني سلبية حمراء و معاني ايجابية زرقاء و اخري ذات رقم 0 اي متعادلة , ويمكن استخلاص المحصلة النهائية للجملة انها كانت ايجابية برقم كذا كذا

مع التأكيد ان النصف الاول فقط من الجملة There are slow and repetitive parts ذات معني سلبي واضح , ولكن ربطها بالجزء الثاني قام بمحو المعني السلبي و اعطاها معني ايجابي

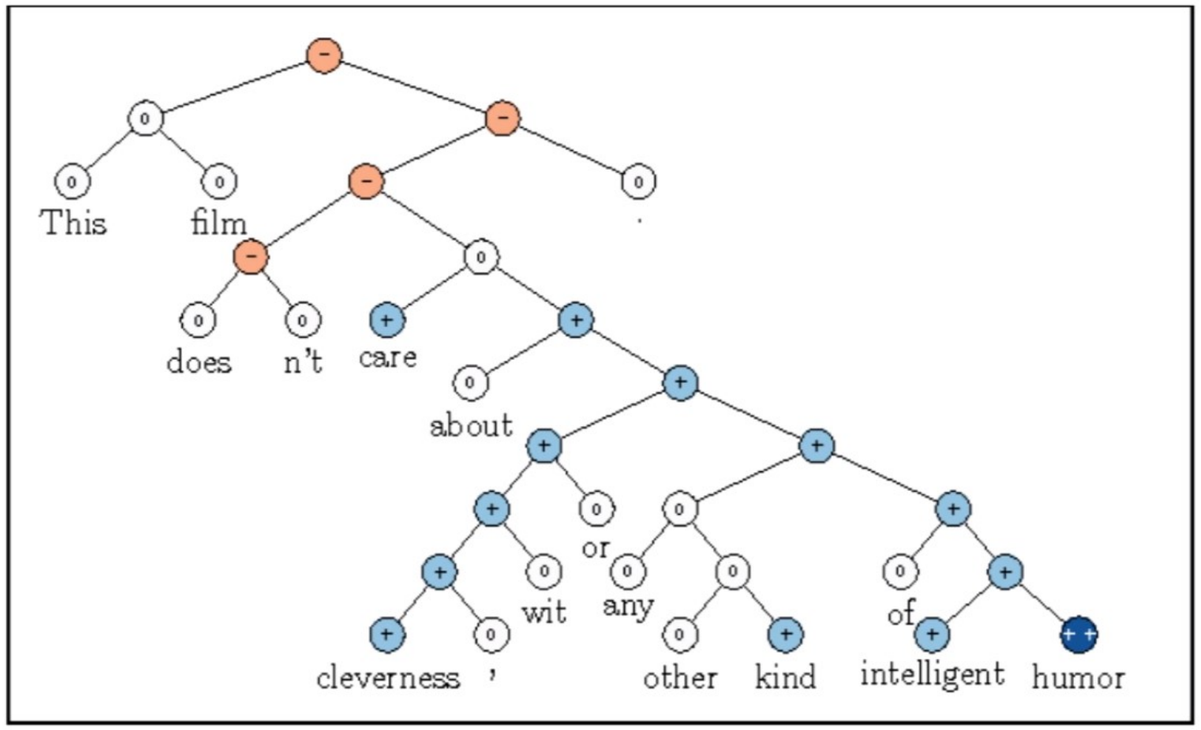
مثل قصيدة الذبياني :

و لا عيب فيهم غير أن سيوفهم , بهن فلول من قراع الكتائب

و يجب أن نعلم ان المحصلة ليس مجرد جمع القيم السالبة و الموجبة معا , اي انها لا تعتمد علي جود كلمات ايجابية اكثر من السلبية , بل علي المعني العام
فهنا مثلا :

This film does not care about cleverness , wit or any other kind of intelligent humor .

فكلمة not قامت بعكس المعني تماما , وقامت الشبكة بفهم ان جميع الكلمات الايجابية ليست في محلها , وانها مفتقدة , و بالتالي فالمحصلة النهائية ان التعليق سلبي , علي الرغم من ان هناك عدد كبير من الكلمات الايجابية و كلمة واحدة سلبية



و لا تنس ان هناك استخدامات مختلفة لكل من RNN و TNN , فالـ RNN يمكن استخدامها لاستنتاج الكلمة الناقصة التالية مثل تطبيقات text generation او question answering

, بينما TNN لا يوجد لديها كلمة تالية , لان الكلمات في تصميم شجري بالفعل , لذا فيمكنها عمل sentimental analysis للجمل , و تصنيفها و معرفة هي تابعة لأي صنف .

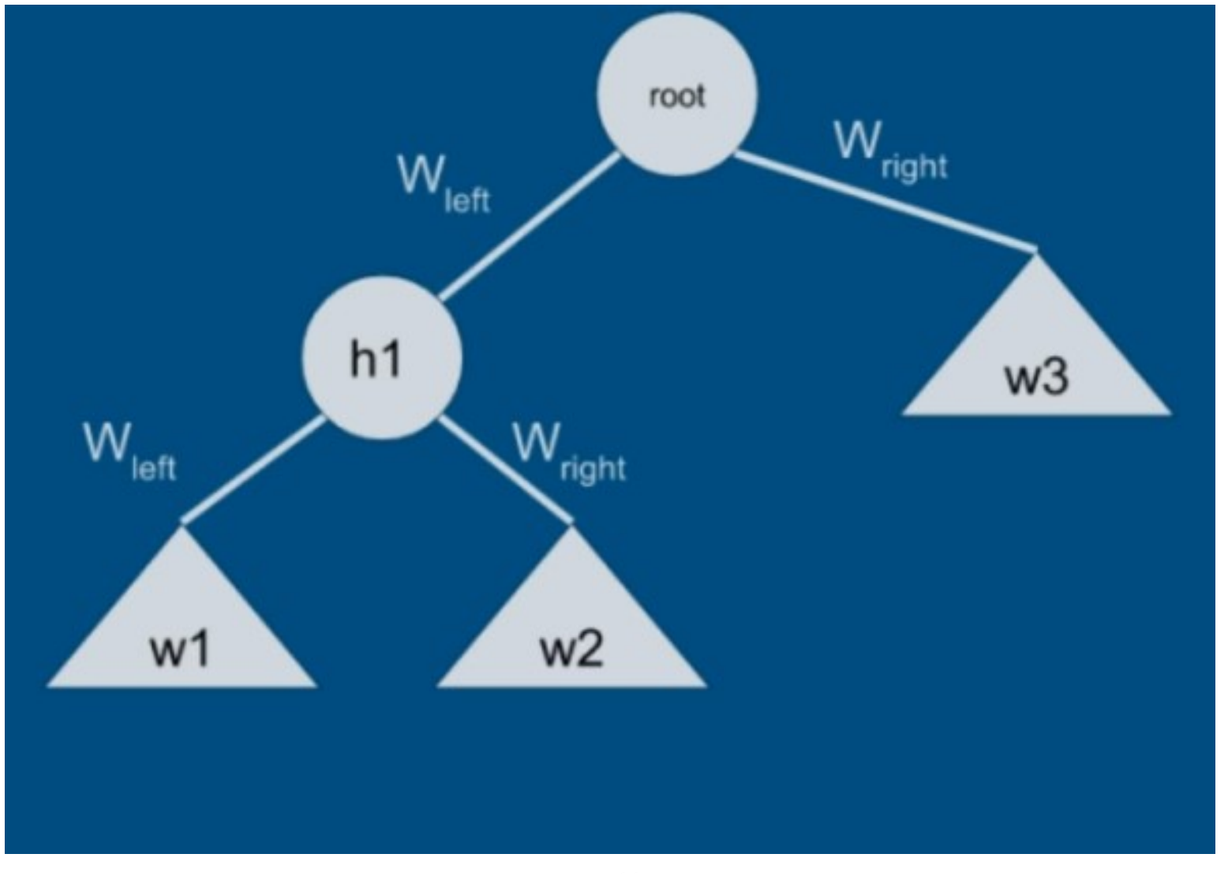
و يتميز TNN علي باقي انواع الخوارزميات الأخري , أن باقي الخوارزميات قد لا تنتبه الي كلمة not التي قد تقوم بتغيير المعني بالكامل لجملة ما , بينما TNN لها قدرة كبيرة علي الانتباه لها و فهمها . . .

ماذا عن المعادلة الرياضية لها ؟

دائما يتم حساب قيمة الخلية اعتمادا علي قيم الخلايا التابعين لها children nodes

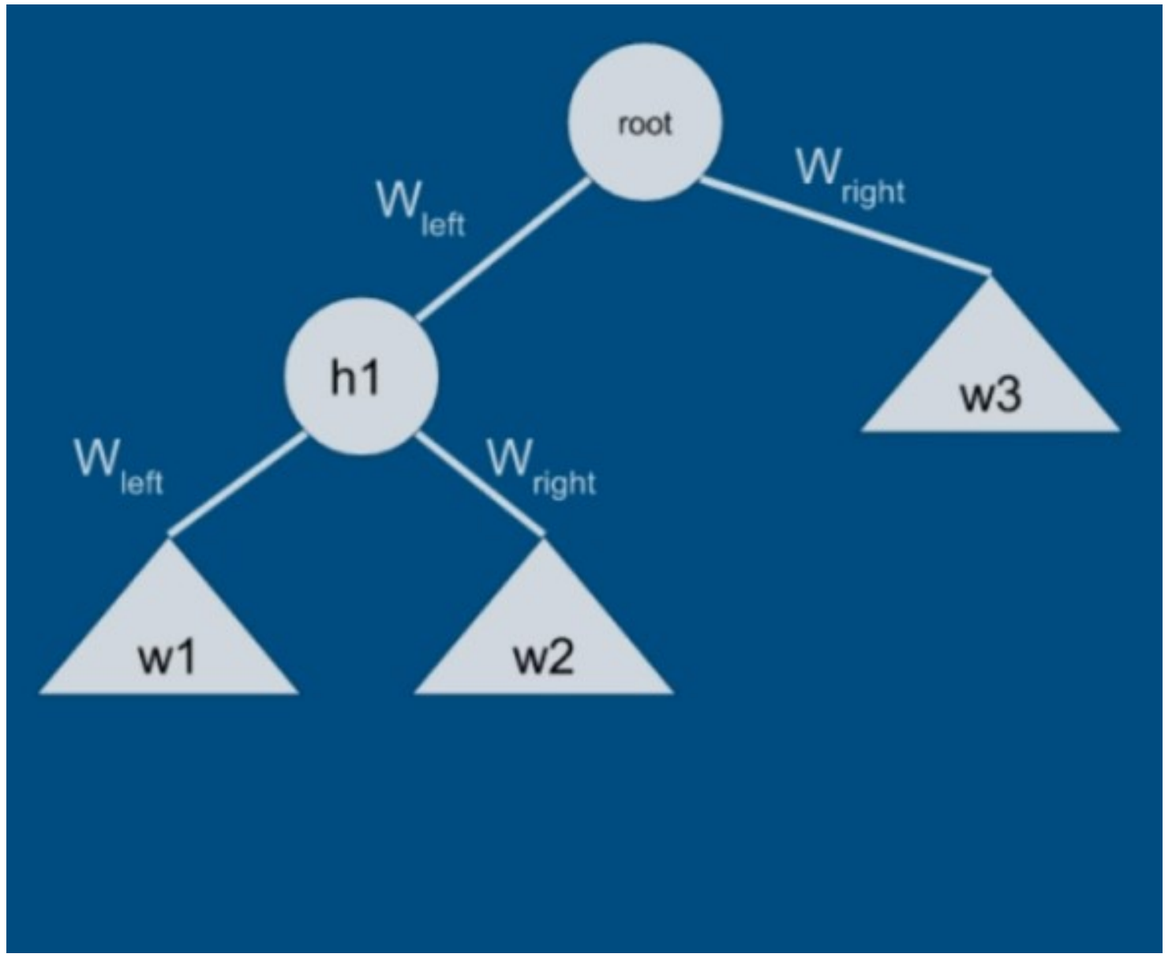
فالخلية h1 تكون معادلتها

$$h_1 = f(W_{left}x_1 + W_{right}x_2 + b)$$



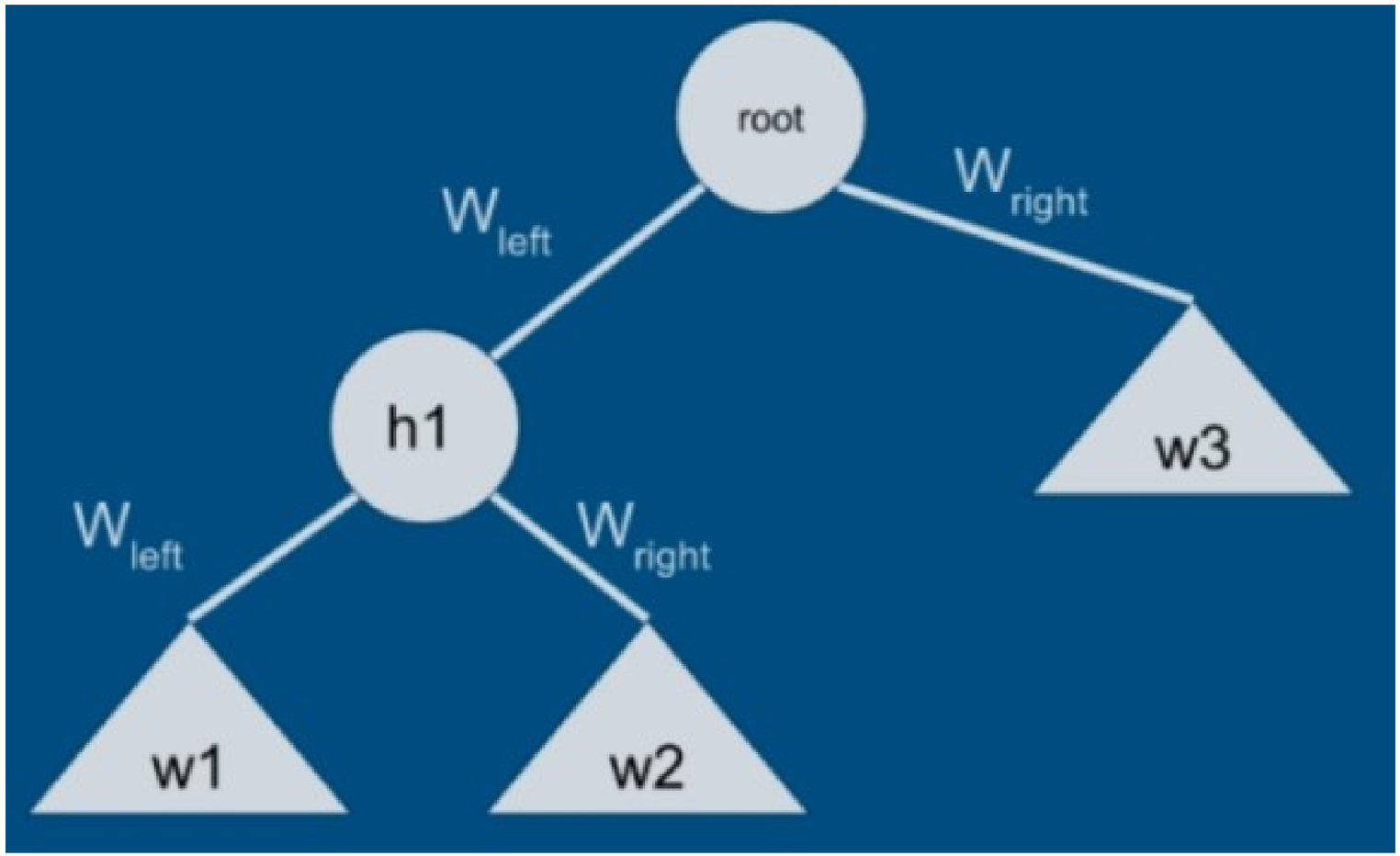
و الخلايا الأعلي تعتمد علي قيم الخلايا الاسفل منها

$$h_{root} = f(W_{left}h_1 + W_{right}x_3 + b)$$



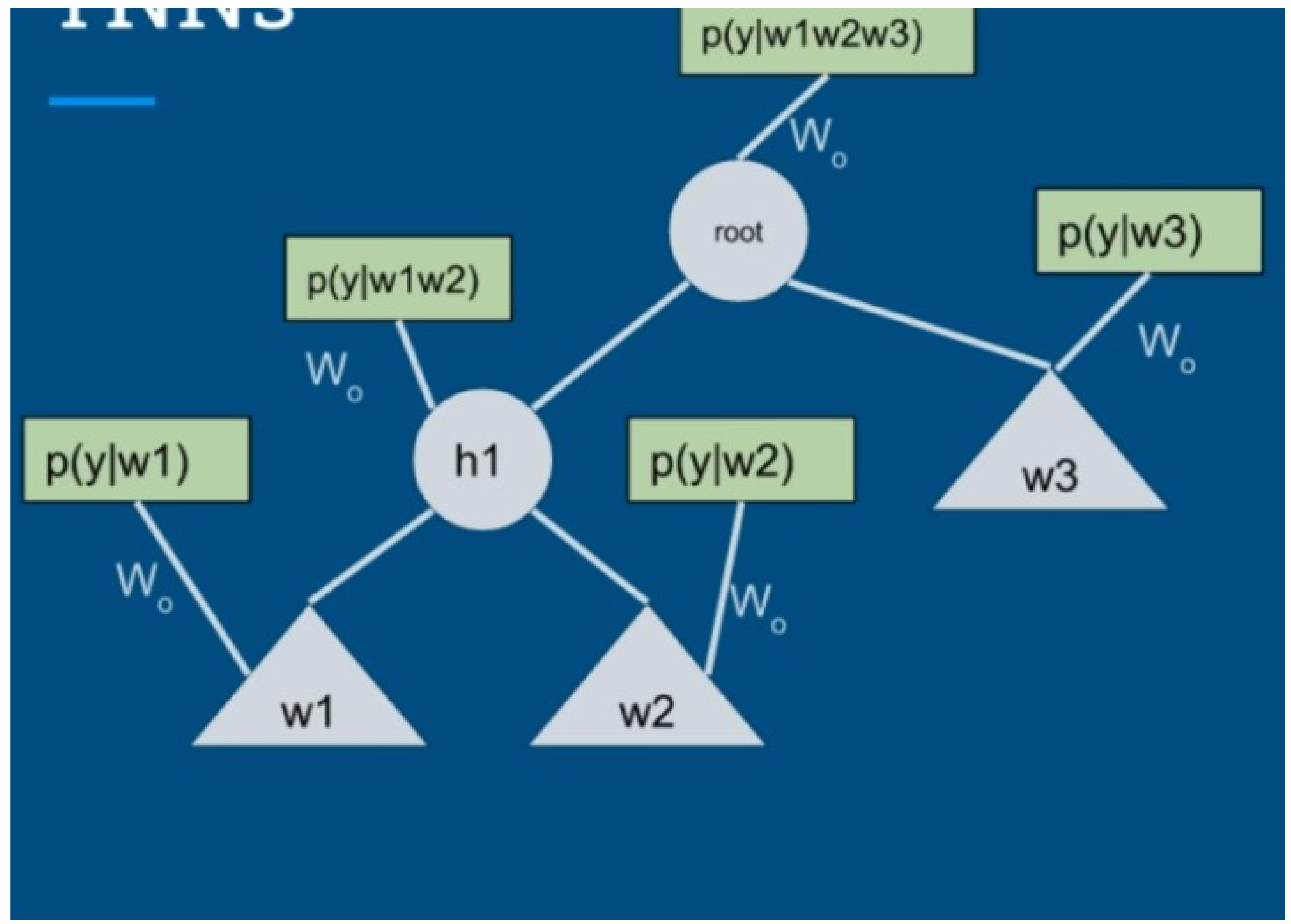
علي ان تكون W هي قيمة مصفوفة الاوزان و قيمة b معامل الانحراف , اما الدالة فهي احد دوال التفعيل

$$h_{root} = f(W_{left}h_1 + W_{right}x_3 + b)$$



و في النهاية يتم تطبيق ال softmax في النهاية لتحديد القيمة النهائية للجملة هل هي ايجابية ام سلبية (او من اي صنف اذا كانت تصنيف متعدد)

$$p(y|h) = \text{softmax}(W_o h + b_o)$$



*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*