International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 1, ISSN No. 2455-2143, Pages 141-147
Published Online May 2021 in IJEAST (http://www.ijeast.com)

CONCEPT OF SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES

Shuchita Mudgil, II year M.tech
Kalaniketan Polyetchnic College Jabalpur

Prof,Ashok Verma, HOD (CSE)
Gyan ganga Institute of Technology and Science

Abstract - Sentiment analysis is used to conclude the approach of a consumer with respect to some topic. Sentimental analysis, a sub discipline within data mining and computational linguistics, refers to the methodology for mining, understanding the opinions expressed by the consumer in various forms like forums, forms blogs etc.
The goal of sentiment analysis is to identify emotional states in online text. We Know human's learns from past knowledge and machines follows instructions given by humans. But what if humans can prepare the machines from the past data and to put output to work much faster well that what is machine learning is it's not about learning it's also about understanding. So we will learn about analysis of sentiments using machine learning techniques

Keywords: Sentiment analysis, machine learning

I. INTRODUCTION

Sentiment analysis is the procedure of calculating, identifying and grouping views represented in a form of text, specifically in order to identify whether the authors behavior towards a particular task is positive, neutral or negative. Opinion Mining also refers to NLP (Natural Language Processing), biometrics, text analysis and computational linguistics in order to detect, extract and refer subjective information. Sentiment analysis basically aims to identify the attitude of a writer with respect to a topic or the complete polarity to a document. The behavior may be a valuation or judgmental or affective state of the author or the emotional communication or interlocutor. It is the calculative study of users opinions, views, behavior and emotions toward an object. Sentiment mining helps to gather positive, negative or neutral information about a product. Then, the highly counted opinions about a product are passed to the user. For promoting marketing, big companies and business magnets are making use of this opinion mining.

Using given studies by Behdenna , et al [1], sentiment analysis is being performed at three levels i.e.:

- **Document level analysis**: The task at this level is to determine the overall opinion of the document. Sentiment analysis at document level assumes that each document expresses opinions on a single entity.
- **Sentence level analysis**: The task at this level is to determine if each sentence has expressed an opinion. This level distinguishes the objective sentences expressing factual information and subjective sentences expressing opinions. In this case, treatments are two fold; firstly identify if the sentence has expressed or not an opinion, then assess the polarity of opinion. But the main difficulty comes from the fact that objective sentences can be carrying opinion.

- **Aspect level analysis**: This level performs a finer analysis and requires the use of natural language processing. In this level, opinion is characterized by a polarity and a target of opinion. In this case, treatments are twofold: first identify the entity and aspects of the entity in question, and then assess the opinion on each aspect

Sentiment Mining has become extremely popular in the field of research technique. A lot of research has already been done but still certain challenges to sentiment mining still exist related to unstructured data. As per the study of various published papers, it can be accomplished that supervised techniques provide much better accurate result in comparison to dictionary technique.

II. LITERATURE REVIEW:

Types of sentiment analysis [16]

1. Manual processing: Human interpretation of the sentiment must be accurate.

2. Keyword processing: Assign positivity or negativity to individual words and calculates the overall percentage score to the post.

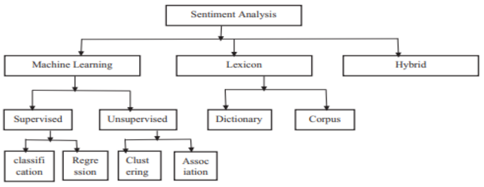
3. Natural language processing (NLP): Also called text analytics, computational linguistics.

NLP is superior to keyword processing. NLP works by analyzing language for its meaning. The information what the vendors get from sentiment analysis provides them to improve their marketing strategy. By sentiment analysis, the researcher can see the positive or negative discussions among their audience. By sentiment analysis, the researcher know the customer's opinions about their views. The opinion are not judged by their functionality, instead of how well it is presented on the online reviews. Sentiment analysis can be measured using They are

- ➢ machine learning,
- ➢ lexicon based, and
- ➢ hybrid-based approaches.

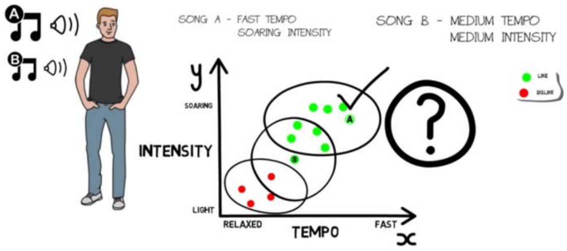
In the **machine learning approach**, the supervised learning model can be easily trained, and the unsupervised model can be easily categorized the data. The **lexicon-based approach** can be easily calculate the sentiment scores for each word. The **hybrid** is a combination of both machine learning and lexicon-based approaches and measures the sentiment for noisy and less sensitive data. [16]

The sentiment analysis can be divided into different categories as shown in Figure



We Know human's learns from past knowledge and machines follows instructions given by humans. But what if humans can prepare the machines from the past data and to put output to work much faster ,well that what is machine learning is ,it's not about learning it's also about understanding. So we will learn about fundamentals of machine learning.

So let's take example of a person named Shuchita , she loves listening to songs ,she either likes them or dislikes them, and she decides them on the basis of Tempo , variety ,intensity and the gender of voice for simplicity lets use intensity and tempo now



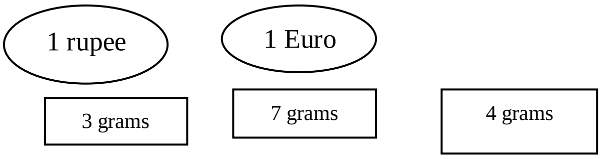
- ➢ Here Tempo is on the X axis ranging from relaxed to fast.
- ➢ Intensity is on the y axis ranging from light to soaring
- ➢ Chetan likes the song with fast tempo and soaring intensity while he dislikes the song with relaxed tempo and light intensity.
- ➢ So now we know Chetan choices ,now let's see Chetan listens to a new song. Lets name it as song A , song A has fast tempo and soaring intensity , so it lies somewhere here looking at the data chart ,You guess where the Chetan will like the song or not ,correct so Chetan likes the song by looking at Chetan's past choices we are able to classify the unknown song very easily right let's say now.
- ➢ Chetan listens to a new song ,lets label it as song B , so song B lies somewhere here, with medium tempo. And medium intensity neither relaxed , nor fast ,neither light nor soaring now can you guess whether the Chetan likes it or not.
- ➢ Not able to guess with this Chetan will like it or dislike it other choice is unclear correct, we could easily classify song A but when the choice become complicated as in the case of song B yes and that's where machine learning comes in , lets see how in the same example for song B if we draw a circle around the song B we see that there are three votes for like where as one vote for dislike , if we go for the majority words, we can say that Chetan will definitely like the song that's all this was a basic machine learning algorithm also , its called K – nearest neighbors, so this is just a small example in one of the many machine learning algorithm .
- ➢ Quite easy right believe it is , but what happens when the choices become complicated as in case of song B that's when machine learning comes in it learns the data , builds the prediction model and when the new data point comes in it can easily project for it more the data better the model higher will be the accuracy there are many ways in which the machine learns it could be either :

  - ✓ Supervised learning
  - ✓ Unsupervised Learning
  - ✓ Reinforcement Learning

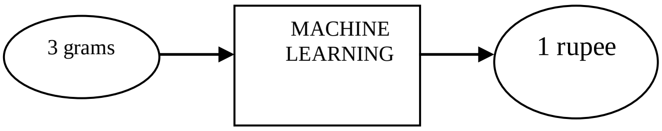
Lets first quickly understand:

Supervised Learning:-

Suppose a friends gives you 1 million coins of three different currencies, say one rupee , one euro , and one dirham , each coin has different weights for example a coin of one rupee weights 3 (three) grams , one euro weighs 7(seven)grams , and one dirham weighs 4 (four) grams



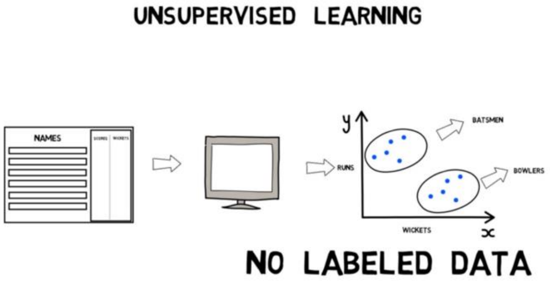
Your model will predict the currency of the coin , here your **weight** becomes the **feature** of the coins while **currency** becomes the **label** when you feed this data to the machine machine learning model , it learns which feature is associated with which label , for example it will learn that if a coin is of three grams ,it will be a one rupee coin . Let's give a new coin to the machine on the basis of the weight of the new coin your model will predict the currency .Hence supervised learning uses labeled data to train the model here the machine knew the features of the object and also the labels associated with those features



On this note let's see the difference with unsupervised learning

Unsupervised Learning:-

Suppose you have cricket data set of various players with their respective scores and the wickets taken, when you feed this data set to the machine , the machine identifies the pattern of player performance so it plots this data with respective wickets on the X axis while score on the Y axis while looking at the data you will clearly see that there are two clusters ,the one clusters are the players who scored high scores and took less wickets

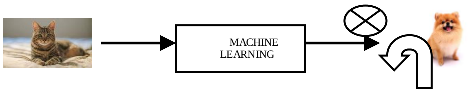


While the other clusters is of the players who scored less runs but took many wickets, so here we interpret these two clusters as batsman and bowlers . The important point to note here is that there were no labels of batsman and bowler, hence the learning with unlabeled data is unsupervised learning . So we saw a supervised learning where the data was labeled and the unsupervised learning where the data was unlabelled .

And then there is :

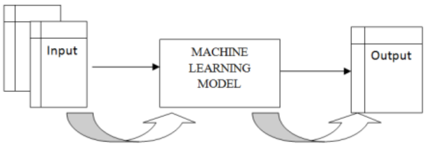
Reinforcement learning:

Which is reward based learning or we can say that it works on the *principle of feedback* here let's say you provide the system with an image of cat and ask it to identify it , the system identifies it as a dog so you give a negative feedback to the machine saying that it's a cat's image , the machine will learn from the feedback



And finally if it comes across any other image of cat it will be able to classify it correctly that is reinforcement learning.

To generalize machine learning model let's see a flowchart input is given to machine learning model which then given the output according to the algorithm applied



If its right we take the output as a final result ,else we provide feedback to the train model and ask it to predict until it learns .

Don't you sometimes wonder how is machine learning possible in today's era , well that's because today we have humongous data available. Everybody is online either making a transaction or just surfing the internet and generating a huge amount of data every minute and that data my friend is the key to analysis also the memory handling capabilities of computers have largely increased which helps them to process such a huge amount of data at hand without any delay and yes computers now have great computational powers, so there are a lot of applications of machine learning out there , to name a few machine learning is used in :

- ❖ **Healthcare** :- To diagnostics are predicted for doctors review
- ❖ **Fraud Detection** :- in the finance sector
- ❖ **Sentiment Analysis** :-

That the technology grants are doing on social media is another interesting application of machine learning

❖ **E-Commerce:-** Also to predict customer churning in the e-commerce sector

While booking a cab you must have encountered a surge pricing often where it says the far row field trip has been updated continue booking yes please .I am getting late for office ,well that's an interesting machine learning model which is used by global taxi giant OLA ,UBER and others where they have differential pricing in real time based on :-

➢ Demand
➢ Number of cars available
➢ Weather
➢ Rush hours etc..

So they use the surge pricing model to ensure that those who need a cab can get one also it uses predictive modeling to predict where the demand will be high with the goal that drivers can take care of the demand and surge price can be minimized.

"Hey Alexa can you remind me to book a can at 6pm today:

Alexa: Ok I will remind you "

Thanks There are many interesting everyday examples around us where machines are learning and doing amazing jobs .

Conclusion of Sentiment Analysis :

Reaching your pockets hopefully your phones still there , our phones do a lot for us ,they check the weather, they remind us to turn on our alarm, just in case we don't wake up the next morning . But there's one thing our phones can't do yet tells how we are, hey "Alexa " hey "Siri" how am I doing today , see these seem like ridiculous questions but with advancements in sentiment analysis and machine learning , our machines are becoming closer to answering these very questions , let me give you a sentence "I loved that movie" , and I asked you to rate it out of the 10 . With 0 being negative and 10 being positive now .

We did all agree that this is pretty positive sentence and give it around to 10.

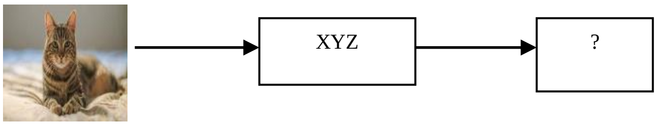
Lets change the verb a bit " I liked that movie" , here still pretty positive , but clearly lower on the scale , now lets go to the other other end of the spectrum , " I hated that movie" , now whoever said this clearly feels negatively about the subject and so we did probably give this around a zero.

Now sentiment analysis is simply using machine learning to teach computers to do just this extract the sentiments out of our sentences. Now does this work? What is machine learning, it's simply a function in math you give it one or many numbers and it spits out another in machine learning these functions are called <u>models.</u>

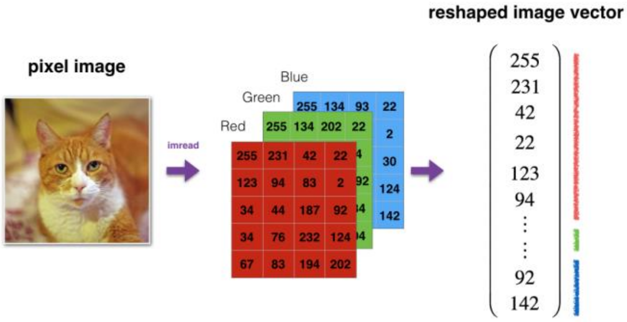
Now these models are often neural networks that simulate the structures of our brains , to set to get inputs and their associations to build models predicting future inputs.

Now here XYZ ,who's XYZ you might ask ?

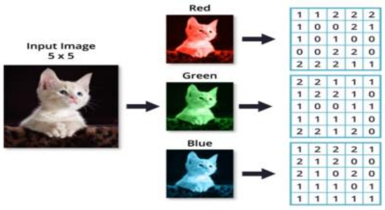
XYZ is our friendly neighborhood machine learning model of course , Now we want XYZ to tell us whether or not this image is a cat ?



➢ XYZ is pretty sad right now because he has no clue what to do ? Here is where we train our model , in order for XYZ to tell us what a cat is and what a cat is not . We as humans must need to first tell , what a cat looks like and what doesn't , there is slight problem here however XYZ doesn't see the image like we do , the one thing XYZ can do however is interpret number , So what we can do is give these images to XYZ , as a list of numbers of RGB vectors for each pixel



➢ Now we break that down RGB vectors are red , green ,blue vectors , for each pixel denoting the color of each pixel in an image .



➢ Now what that effectively allows us to do is to convert these images into numbers , that's great XYZ can now understand what we are trying to give XYZ , he knows what he's trying to do , now we can do when given an unfamiliar image , is then turn into RGB vectors give it to XYZ , and hey what do you know he thinks it's a cat he is 97% sure actually , so this was the case with pictures with images but we are talking about sentences ,the thing is its exactly the same how do we turn words into numbers , now some of you might be thinking lets slap a number on each word and call it a day , but the thing is if we train our models using those vector inputs, we did run into a problem this method struggles to recognize similarity between a word such as loved and liked as opposed to a low similarity between loved and hated . here is where we run into one of the most fundamental concepts in sentimental analysis.

- Word vectors now, what are verb vectors well they're exactly as they would seem, they are vectors corresponding to each word, much like the RGB vector for each pixel.

- Now unlike the RGB vectors however these word vectors can span from 25 upto thousand components now conveniently as these vectors are still simply a list of numbers they can be plotted on an n dimensional space, but for the sake of visualization and your brains, lets reduce that down to two on this coordinate plane. What word vectors allow us to do is to demonstrate and evaluate the relationships between words as distances between points, now somewhere on this coordinate plane. Lion and cat would be near to each other related to their resemblance.

- Somewhere in the middle we have no clue , we have run into the dilemma which makes it possible for word vectors to be multidimensional , by adding more vectors more dimensions to these vectors , we are able to express the relationship between words in the English language with more fine distinction, great now we have these word vectors , we can associate them to the words in our sentence , converting them to numbers that XYZ loves theoretically now , we can feed these number to XYZ and XYZ will now be able to predict the sentiment of any sentence we give it , So naturally I decided to put that to test .

A Machine Learning model

I loved that movie	I hated that movie
1	0

- Where would I get my data to train this model well after some searching I decided to go with Kaggle's twitter sentiment data set consisting of millions of tweets manually categorized by either zero for negative or one for positive. But just as you and I will be better at identifying something the more examples you got of it :-
- XYZ can benefit from as much data as we can give it as for our word vectors, however when I went with study of standford university glove stand for global vectors , now this word vector st was pre created , which means that these researchers had to go through thousands and thousands of sentences look at instances for each word and evaluates their context to create word vectors for each and every word , there was one more step I had to take before I could train XYZ, however lets look at this tweet

@username Stopped at KFC for lunch!so excited ☺ #nuggets

Now if we fed this right to our model we did see a problem, see us as humans can see through the twitter clutter can see through the various distractions in this tweet but for XYZ , XYZ need a bit of help and that's why we need to clean the data set show you whats I mean the first things to go were punctuation along with punctuation when

Stopped at KFC for lunch so excited

- Twitter artifacts such as mentions Hash tags and links , second went are what recalled in Natural Language Processing (NLP) as stop words , words such as , as if I and that don't necessarily add to the meaning of the sentence , now finally and arguably the most tricky part of cleaning this data set was how to deal with internet slang , now it's impossible to go anywhere on the internet without encountering some sort of abbreviation some sort of slang , the tough part about dealing with this is that there is no set way of evaluating these words, now to be fair common words such as law or lmao all have their individual entries in word vector sets such as glove, now misspellings such as the one we see in this tweet here can be caught with a spell check , but some words and phrases do end up slipping through our fingers and that does make or break some sentiment analysis models,

	100 elements
Stopped	{-0.567,1.367,………-0.6169}
KFC	{-0.395,0.821,………1.212}
Lunch	{-0.251,0.951 ,……….-0.232 }
Excited	{-0.934 ,0.823,………-0.236}

- Now regardless we have caught that and now we were able to condense that original tweet into the four words that you see on the bottom there , now that we have cleaned our tweets , we can associate the word vectors in gloves to each word and now again we have our numbers to word association and can now train our model that's exactly what I did , now how is my model you might be asking how good was it , well luck its for the safety of the internet world as we know it , I was not that successful my model reached around a 60% accuracy which is meant that it was able to correctly identify the sentiments of around 60% of the sentences that I gave :

60% + accuracy

- It however considering that this is a problem that yet to be solved this number is a sign of hope for things to come . Now throughout reading this paper you might have been asking yourself why do we care who asked and the what's next , and I concluded after reviewing the research papers , its true this technology is bringing us ever so closer to our inevitable robot overload world , but I still believe that this technology is very important and essential to our technological development for the benefits that it can provide .

Applications of Sentiment Analysis are purely commercial :

- We see movie producers using sentiment analysis to evaluate audience feedback on their projects.
- Cooperate sectors including this technology to assess how consumers are reacting to their products. Etc..

III. FUTURE WORK

In the future as this technology gets better we can see that this technology gets better we can see that this technology can be applied to a numerous of problems, for example Sentiment analysis can be used to provide help for people with mental health issues , many people with these issues get safe haven in the internet and so with this technology we will be able to provide help for people that might have been reluctant to seek it ..

This Technology can be can be used by government to make the internet safer for all of us and hey if that didn't reach all of you then may be our phones can become a counselor one day , "HEY Alexa /Siri how
I am doing , thankyou"

IV. REFERENCES

[1] Document Level Sentiment Analysis: A survey S. Behdenna, F. Barigou and G. Belalem Department of Computer Science, Faculty of Sciences, University of Oran 1 Ahmed Ben Bella, PB 1524 El M'Naouer, Oran, Algeria (31000), published :2018-03-14,EAI , http://dx.doi.org/10.4108/eai.14-3-2018.154339

[2] A Survey on Sentiment Analysis for Big Data Swati Sharma, Mamta Bansal, Ankur Kaushik.vol no 6,issue no .06, june 2017,International journal of Advanced Research in Science and Engineering

[3] F. R. Lucini et al., ``Text mining approach to predict hospital admissions using early medical records from the emergency department," Int. J. Med. Inf., vol. 100, pp. 1_8, Apr. 2017.

[4] Z. Khan and T. Vorley, ``Big data text analytics: An enabler of knowledge management," J. Knowl. Manage., vol. 21, no. 1, pp. 18_34, 2017.

[5] T. T. Thet, J.-C. Na, and C. S. G. Khoo, ``Aspect-based sentiment analysis of movie reviews on discussion boards," J. Inf. Sci., vol. 36, no. 6, pp. 823_848, 2010.

[6] H. Yu and V. Hatzivassiloglou, ``Towards answering opinion questions:Separating facts from opinions and identifying the polarity of opinion sentences," in Proc. Conf. Empirical Methods Natural Lang. Process.,2003, pp. 129_136.

[7] R. Piryani, D. Madhavi, and V. K. Singh, ``Analytical mapping of opinion mining and sentiment analysis research during 2000_2015," Inf. Process.Manage., vol. 53, no. 1, pp. 122_150, 2017.

[8] E. Cambria and B. White, ``Jumping NLP curves: A review of natural language processing research," IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48_57, May 2014.

[9] Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges,SHAHID SHAYAA et al, ,IEEE Acess ,Received April 24, 2018, accepted June 18, 2018, date of publication June 28, 2018, date of current version July 30, 2018.

[10] BigData Concepts and Architecture Journal of Computer Science and Information Systems, Volume 1. Issue 4, August 2020 ISSN 2535-1451

[11] A Survey on Sentiment Analysis for Big Data Swati Sharma1, Mamta Bansal2, Ankur Kaushik3 IJARSE ,Vol no.6 Issue no.06, june 2017

[12] D. Denyer and D. Tran_eld, ``Producing a systematic review," in The Sage Handbook of Organizational Research Methods. Thousand Oaks, CA, USA: Sage, 2009.

[13] D. Tran_eld, D. Denyer, and P. Smart, ``Towards a methodology for developing evidence-informed management knowledge by means of systematic review," Brit. J. Manage., vol. 14, no. 3, pp. 207_222, 2003.

[14] R. J. Light and D. B. Pillemer, Summing Up: The Science of Reviewing Research. Cambridge, MA, USA: Harvard Univ. Press, 1984.

[15] D. M. Rousseau, J. Manning, and D. Denyer, ``11 evidence in management and organizational science: Assembling the _eld's full weight of scientific knowledge through syntheses," Acad. Manage. Ann., vol. 2, no. 1,pp. 475_515, 2008.

[16] B. Usharani ,"Analysis of Supervised and Unsupervised Learning Classifiers for Online Sentiment
Analysis "Asian Journal of Computer Science Engineering 2018;3(4):17-21

[17]. Tsytsarau M, Palpanas T. Survey on mining subjective data on the web. Data Min Knowl
Discov 2012;24:478-514.

[18]. Liu B. Sentiment Analysis and Opinion Mining. San Rafael: Morgan and Claypool Publishers; 2012. p. 1-168.

[19]. Pang B, Lee L. Opinion mining and sentiment anlaysis. Found Trends Inf Retr 2008;2:1-135.

[20]. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis.
IEEE Intell Syst 2013;28:15-21.

[21]. Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013;56:82-9.

[22]. Montoyo A, Barco PM, Balahur A. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675-9.

[23]. Li N, Wu DD. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis Support Syst 2010;48:354-68

[24] Balahur A. Sentiment Analysis in Social Media Texts. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2013. p. 120-8

[25]. Nakagawa T, Inui K, Kurohashi S. Dependency Tree Based Sentiment Classification Using CRFs with Hidden Varaibles, Human Language Technologies: The 2010 Annual Conference of the North America Chapter of ACL; 2010. p. 786-94.

[26]. Moilanen K, Pulman S. Sentiment Composition. The Oxford Computational Linguistics Group. Proceedings of RANLP; 2007. p