Seminar on NLP of Social Networks
(Opinion Mining in Web)

Group-16

Sushmita Bhattacharya (133050022)

Deepa Panwar (13305057)

Ravali (133050090)

Introduction to Opinion Mining

- Most of us hold opinions and one's opinions affect others.
- Since one can publish anything on the web, we can see lots of opinions being expressed in the social media.
- In some way, online communications play a role in transforming the opinions of the society to some extent.
- Traditional methods of survey are costly
- Despite its applications, little is understood about how online opinions emerge, diffuse, and gain momentum.
- The goal of Opinion research is to identify emerging societal trends based on views, dispositions, moods, attitudes and expectations of stakeholder groups or the general public

Building blocks of Opinion Mining

Opinion mining broadly involves:

1. Automated topic and
2. Sentiment or Opinion detection

Opinion Formation in the Web in the context of policy making[2]

Opinion Formation Framework in the context of some government action

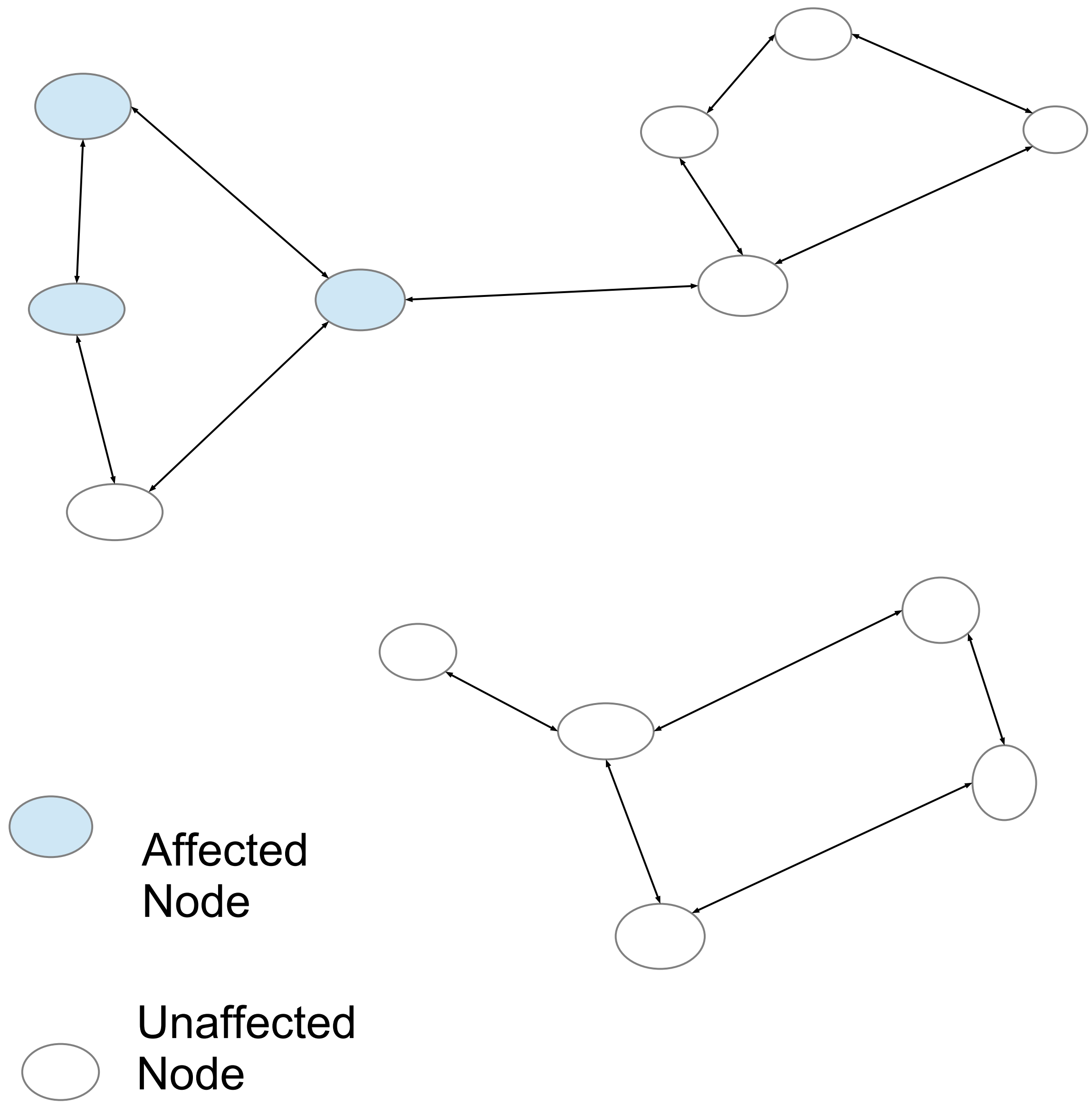
Models of opinion formation considered here has three stages, the initial state, the alert state, and the percolated state. (These models are based on work in sociology and physics)

Lets look at this in the context of reaction to government action.

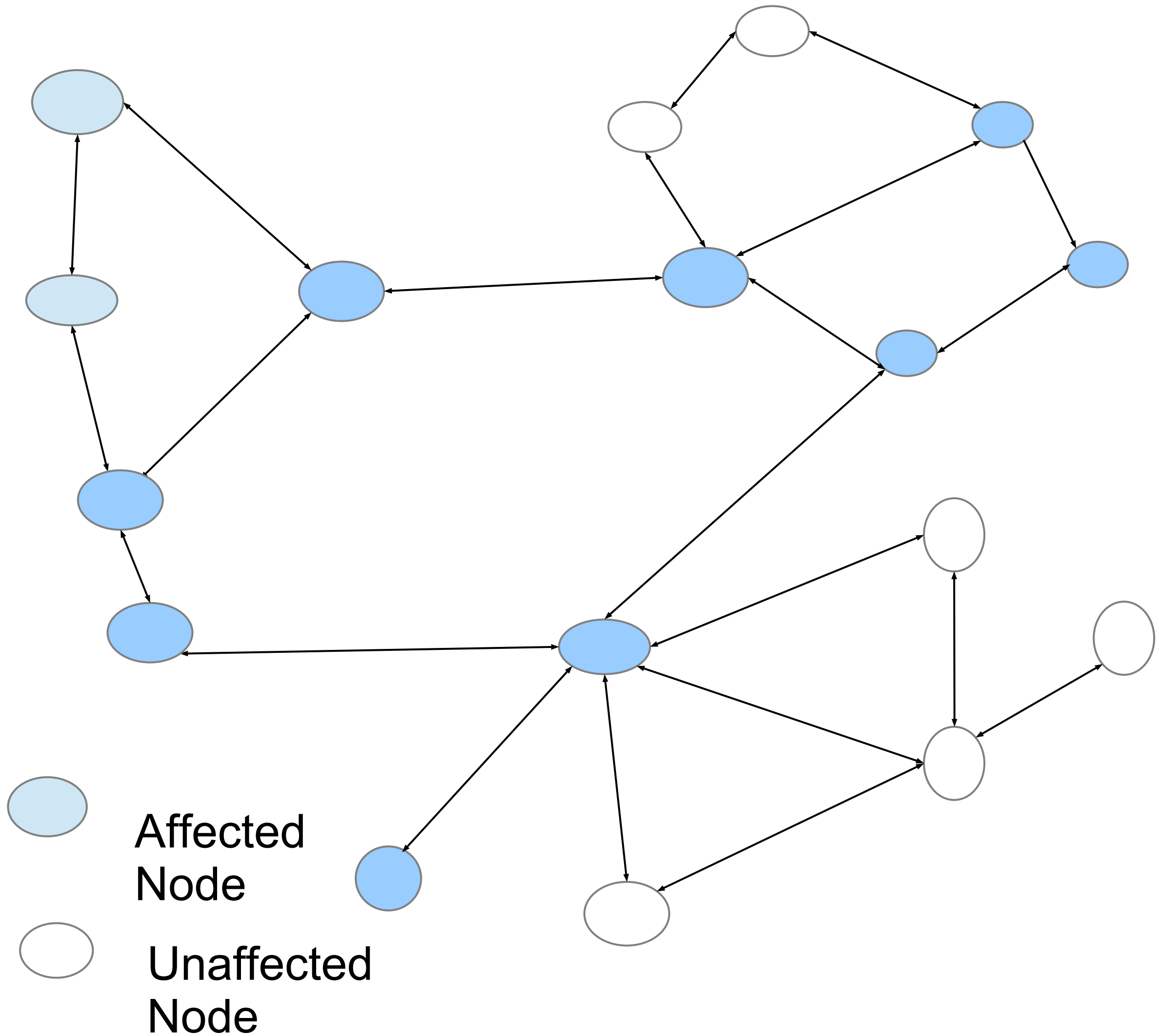
1) The first stage is the **“initial state”** in which the vast majority of participants ignores government action (or inaction) on a potentially important issue.

2) When new information comes, it changes the initial to an **“alert state”**.

Simple model of opinion formation in the context of policy making - Initial State



Simple model of opinion formation in the context of policy making - Alert State

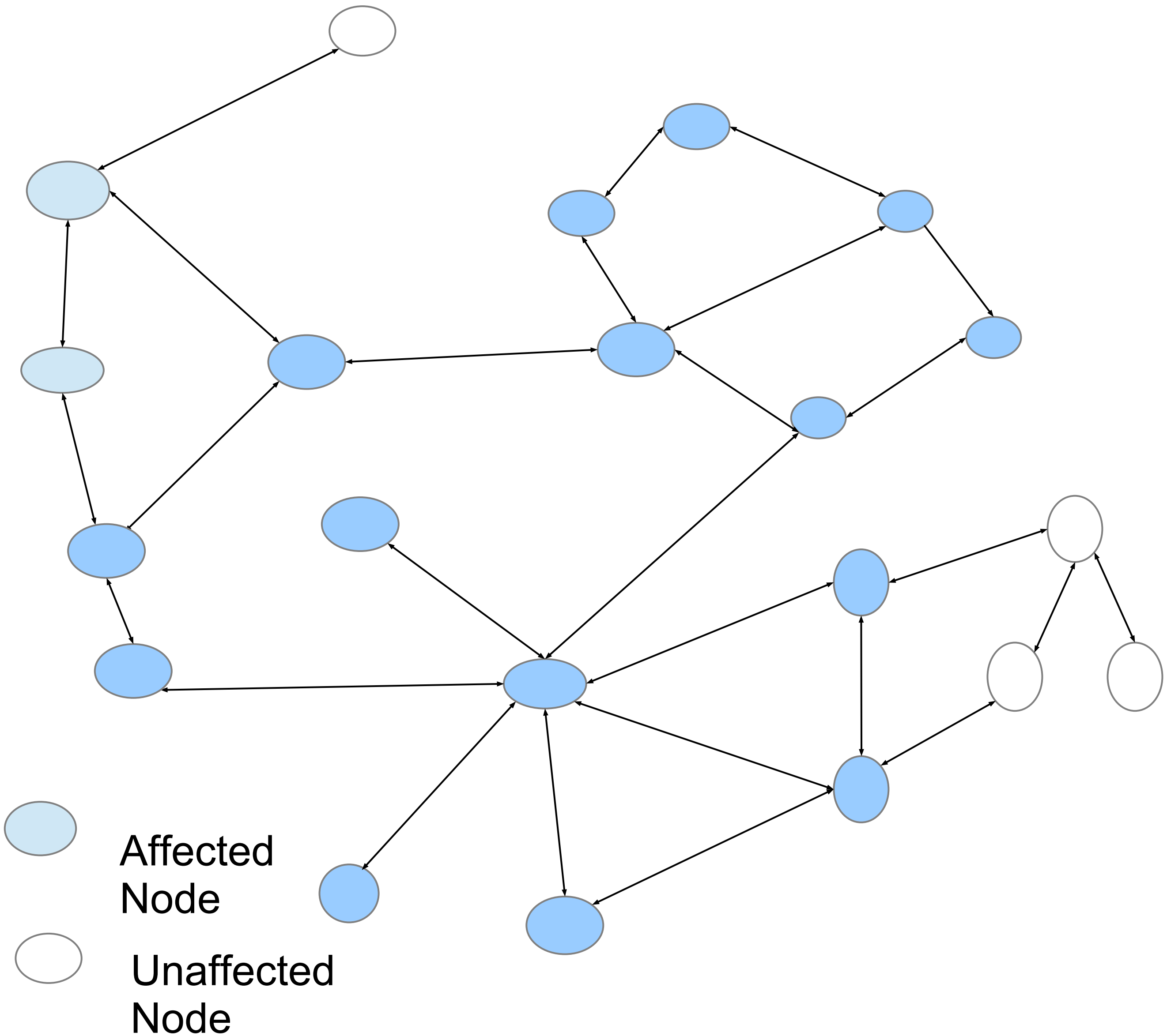


Opinion Formation Framework cont.

3) The arrival of more new information changes the alert state either in the direction of the initial state or towards a "**percolated state**".

If this information is perceived as reinforcing the earlier assessment, we can see some kind of dynamic motion in the model. With increasingly more intensive interactions, thresholds for other participants adopting the opinion are more quickly reached and the opinions and the corresponding actions percolate.

Simple model of opinion formation in the context of policy making - Percolated State



Few considerations of this model which makes it work

The model takes full account of the reflexivity of agents, because participants adjust their behavior in cognizance and reflection of other agents' behavior and the arrival of new information.

- They can enhance links with the ones holding same opinions or
- They may get disconnected from someone having opposite opinion.

So, because of this dynamic methodology, the reinforcements in the communication network is intensified and the initial network characteristics change to a dominant mainstream in extreme cases isolating the minority opinion holders.

Factors which influence the change of opinion

In real society, some links cannot be broken (eg., family, close friends).

There could be a change in opinion due to peer pressure.

A change of opinion could happen because a person is very strong at influencing or a person easily changes opinions !!

Application of this opinion formation model

With the opinion formation framework discussed, coupled with some feedback from user, effective opinion mining can be built achieved.

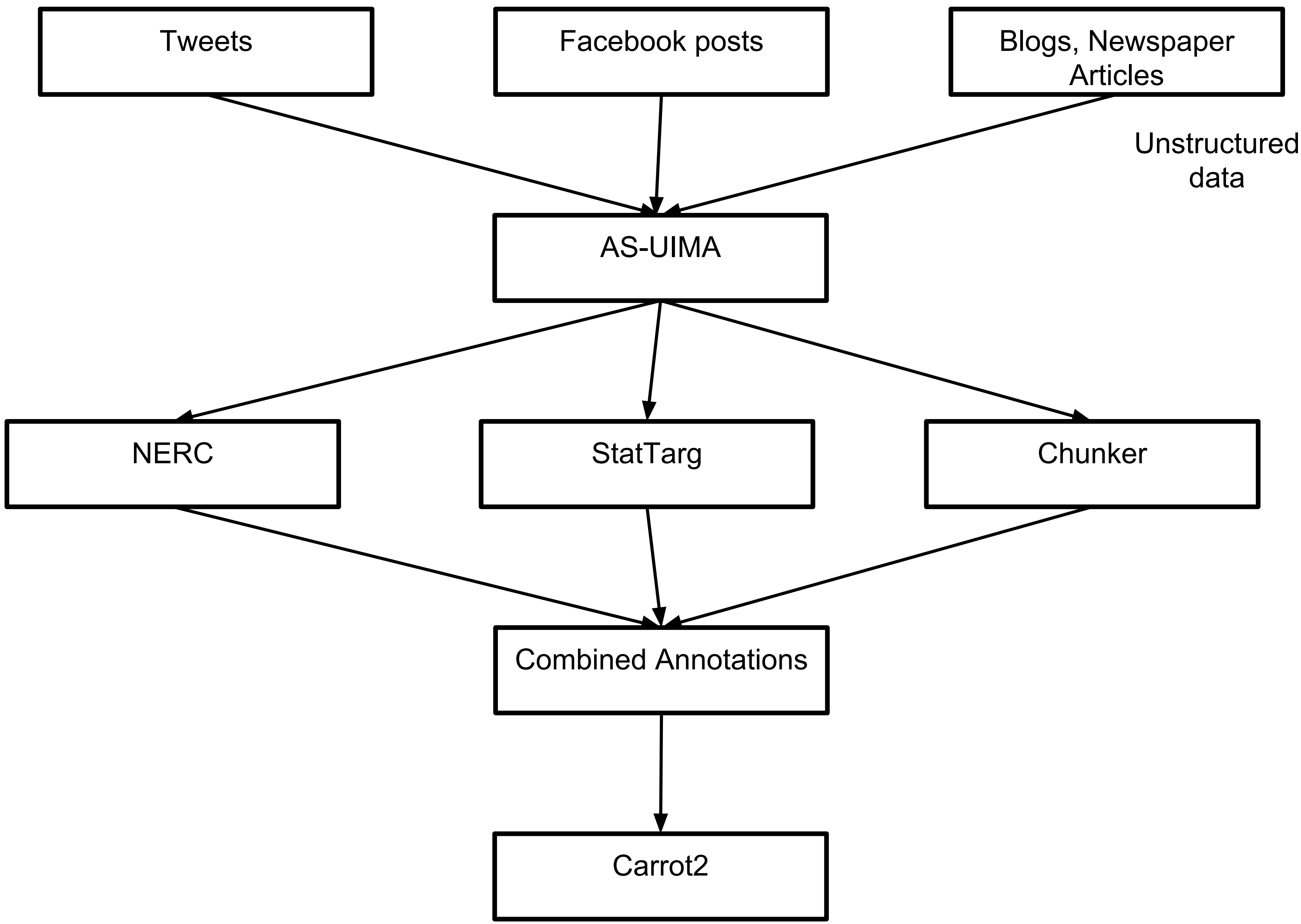
The ultimate goal is to learn from unexpected reactions and the evolution of opinions to bring out accurate decisions and the intended consequences.

A hybrid framework for scalable topic identification Opinion Mining in Social Media[1]:

Various techniques used in the process of topic identification

Introduction and the techniques

- Social media constitute heterogenous information.
- No single technique is capable of interpreting that information.
- So they implemented a hybrid technique for topic identification


Tweets
Facebook posts
Blogs, Newspaper Articles
Unstructured data
AS-UIMA
NERC
StatTarg
Chunker
Combined Annotations
Carrot2

They applied following techniques

1. UIMA(Unstructured Information Management Applications) major goal of UIMA is to transform unstructured information to structured information by arranging analysis engines to detect entities or relations and thus to build the bridge between the unstructured and the structured world.

2.UIMA-AS (Asynchronous Unstructured Information Management Applications) By using UIMA-AS you can run all the annotators parallely in distributed manner and increase the scaleout.

3. Carrot2 is an open source search results clustering engine. It is connected to solr and used to test clustering conditions and algorithms and provide good visualization.

4. Noun Phrase Chunker :- because nouns are more useful . OpenNLP is used as Chunker.

OpenNLP is machine learning based toolkit used for the processing of natural language text.

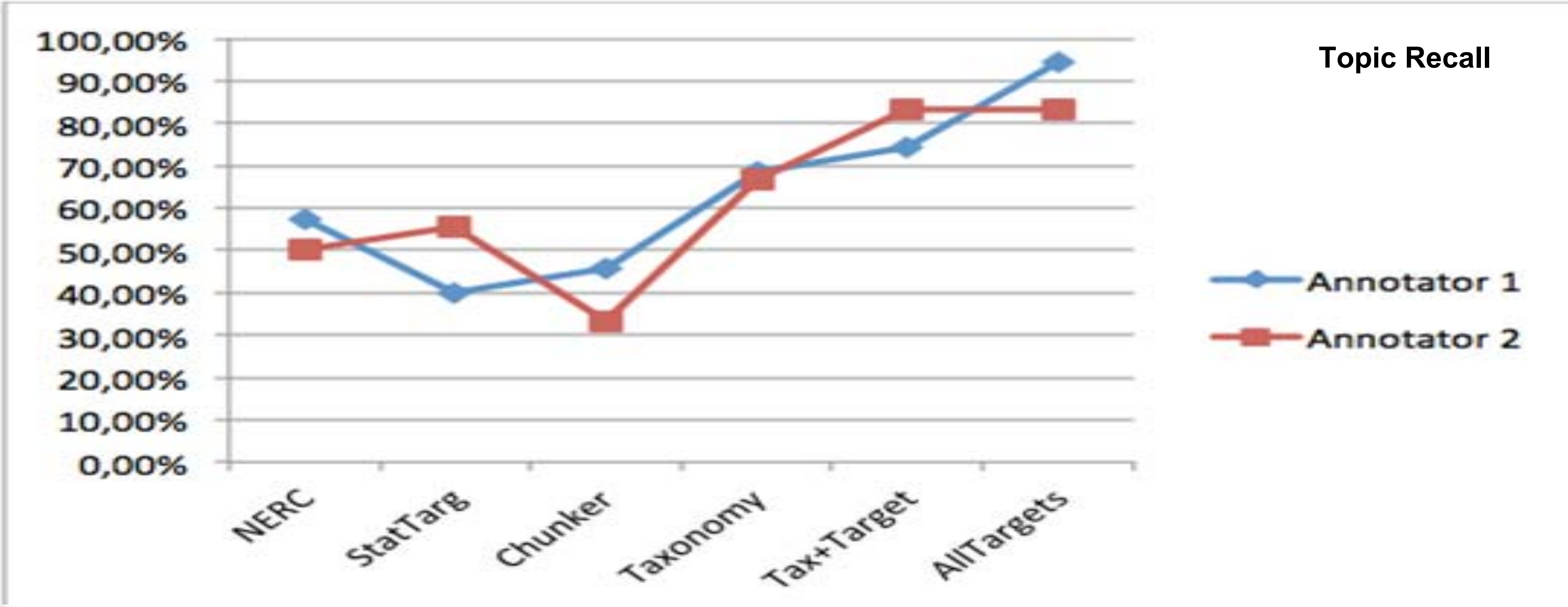
5. Gazetteer Matcher (Taxonomy) or expression matchers:- A gazetteer consists of a set of lists containing names of entities such as cities, organisations, days of the week, etc. These lists are used to find occurrences of these names in text, e.g. for the task of named entity recognition.

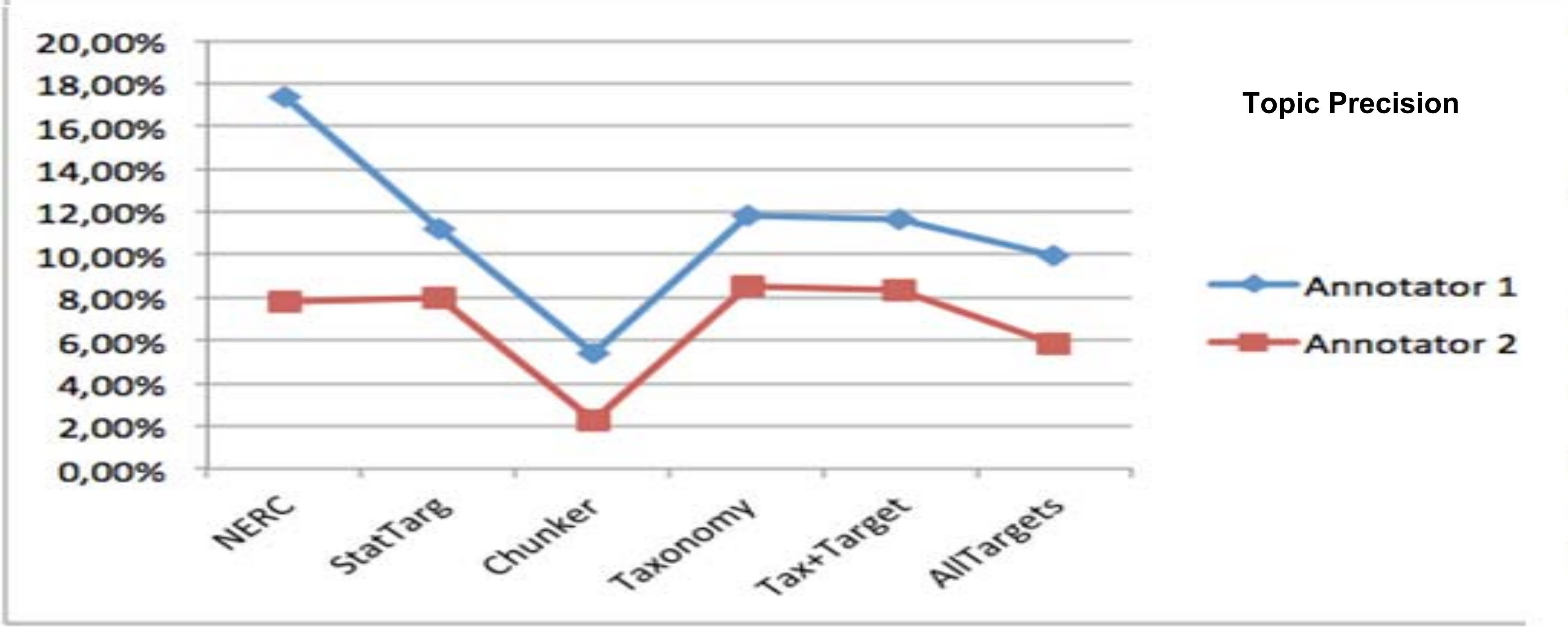
It is used for name entity recognition.

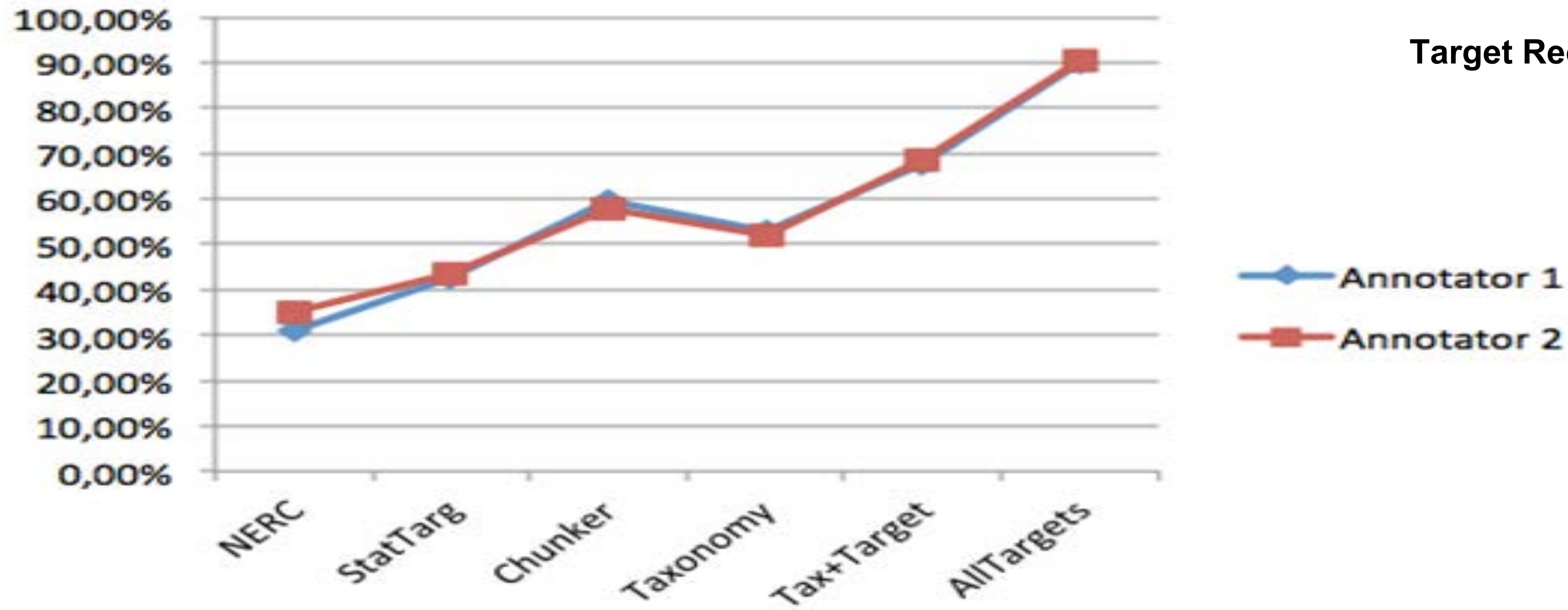
Results of the experiments

A combination of NERC, Chunker and StatTarget had a very high recall of around 90%.

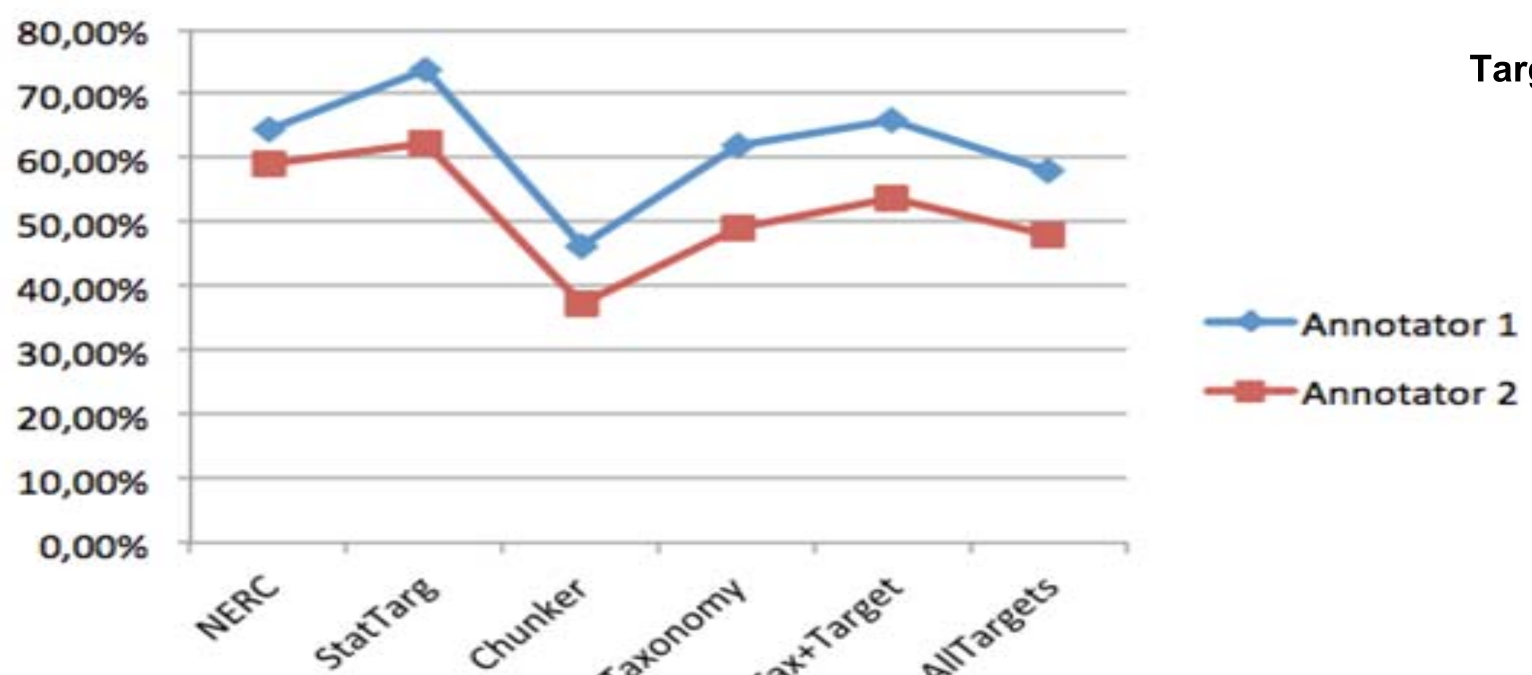
The precision obtained on topic detection is very low because evaluation is done using all candidate on different annotation layers, with no selection process.


Topic Recall
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
NERC
StatTarg
Chunker
Taxonomy
Tax+Target
AllTargets
Annotator 1
Annotator 2


Topic Precision
20,00%
18,00%
16,00%
14,00%
12,00%
10,00%
8,00%
6,00%
4,00%
2,00%
0,00%
NERC
StatTarg
Chunker
Taxonomy
Tax+Target
AllTargets
Annotator 1
Annotator 2



Target Recall



Target Precision

Challenges for processing twitter like online data and the proposed approaches[3]

Micro-blog post is very short and colloquial, sometimes grammatically incorrect.

Traditional opinion mining algorithms like(lexicon based approaches, rule based approaches) do not work well in such type of text

Emoticon based algorithm makes use of “:)”, “:-)” to denote positive sentiment, and “:(”, “:-(” for negative emotional content

Contradictions **Twilight Saga is over :(I'm so sad, I'll miss it”**

We now present an system architecture and algorithm for detecting opinion and their polarity

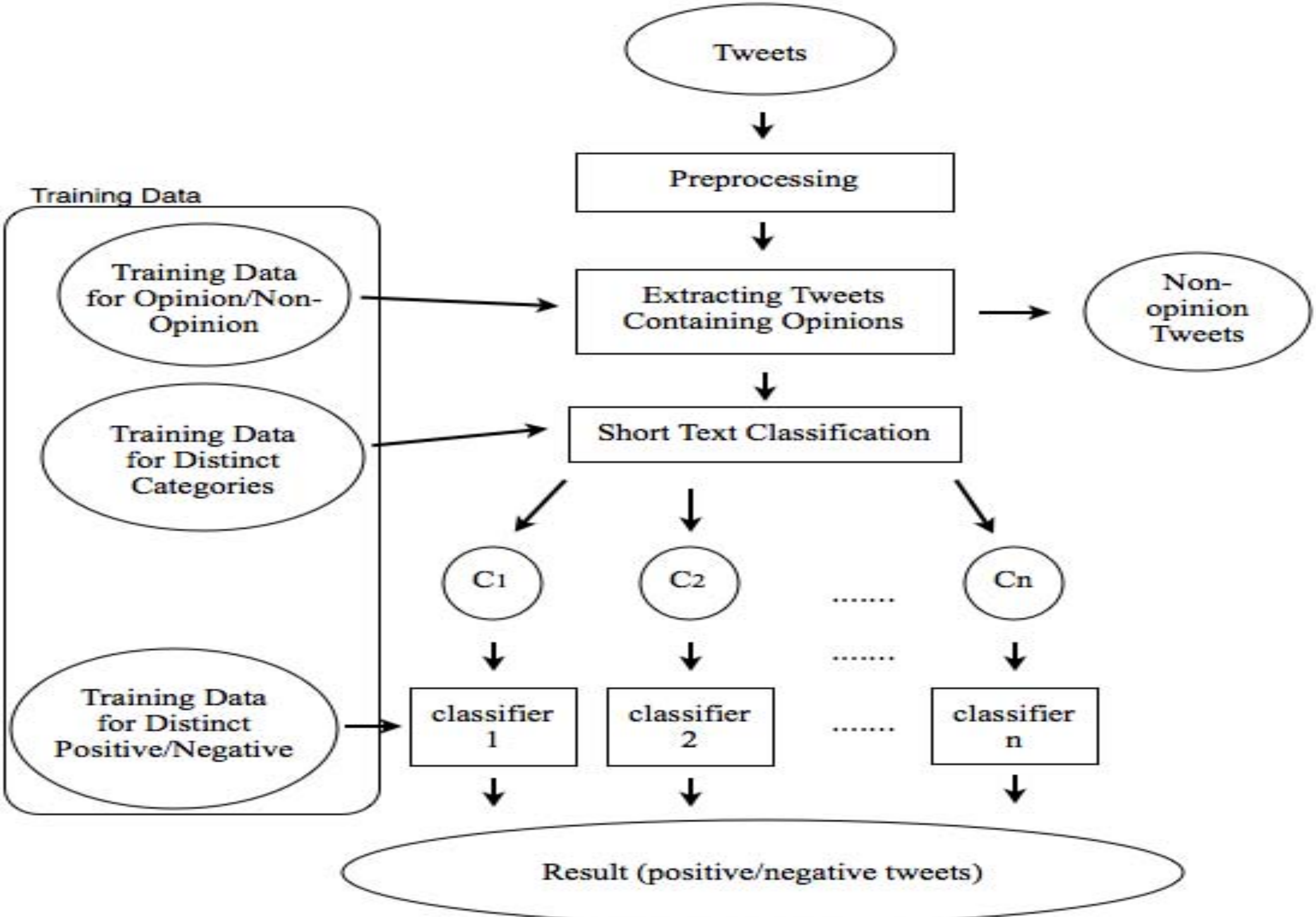
Algorithm

- Naive Bayes (NB) classifier is used in various stages of the algorithm
- Domain-specific training data is used to build a generic classification model from social media data to help improve the performance

Overall flow of the system

- **Data category**: Data from different domain introduce different terms and various meanings. So here data categorisation is done for better accuracy.
- Here small amount of **manually tagged data** used for training. Emoticon based data not used.
- Data crawled from tweets -> preprocessing -> opinion data extracted -> training data categorised into several categories -> Polarity determined.

Flow diagram



Algorithm step 1: Preprocessing

- All tweets converted to lowercase
- Emoticon extracted based on Emoticon dictionary
- @ replaced by USER, #tag used for determining summary.
- POS on learning corpus (tool Tree Tagger)
- English words are extracted by using WordNet (EN/NEN)
- Stop word eliminate(ST/NST) by making use of a stop word dictionary,
- Multichar to single char cooool -> cool
- URL, tweets with very few words or very few word other than greeting words are eliminated
- An acronym dictionary is used for expanding acronym meaning

Algorithm step : Using Naive Bayes

- NB classifier to check training data opinion or not
- We are maximizing the class c given document

$$\tilde{C} = arg\,max\,P(c|d) = arg\,max\,P\,(c)\prod_{1\leq k\leq n} P(t_k|c)$$

t_k = k^{th} token in document d(n=size of document)

P(c) = prior prob for class c

- Calculate $P(t_k|c)$

$$P(tk|c) = \frac{T_{ct}+1}{\sum_t(T_{ct}+1)}$$

Result

DataSet

- Twitter API to collect data
- Crawl tweets of three distinct categories (camera, mobile phone, and movie) from the time period between November 1, 2012 to January 31, 2013.
- English data
- Manually label tweets

Our Algorithm accuracy:

Extracting Tweets Containing Opinions	76.80%
Short Text Classification	96.60%
Training Multiple Classifiers in Distinct Categories	90.17%

Algorithm based on Emoticon based data : 58.65%

Some Improvements can be done

Different Machine Learning techniques can be used in different stages.

Finally, rule-based models or other methods of natural language processing can be incorporated into our system.

Conclusion

- Opinion mining is a type of natural language processing for tracking the mood of the public about various things ranging from consumer product brands to government policy.
- Here Various challenges like Genre/category recognition (e.g. product reviews, political opinions), Content relevancy, overall sentiment extraction etc are addressed.
- Putting together the different techniques and algorithms of Topic identification,sentiment extraction and the opinion formation framework we should be able to identify the opinions for the desired topics on the web. If this is coupled with manual feedback, then accuracy could be pushed further.

References

[1]A hybrid framework for scalable opinion mining in social media: detecting polarities and attitude targets
Authors: Carlos Rodríguez-Penagos, Jens Grivolla, Joan Codina Fibá
Published in: Proceedings of the Workshop on Semantic Analysis in Social Media
Pages 46-52
Association for Computational Linguistics Stroudsburg, PA, USA ©2012
http://dl.acm.org/citation.cfm?id=2389969.2389975&coll=DL&dl=ACM&CFID=372997762&CFTOKEN=43551036

[2]**Opinion mining in social media: modeling, simulating, and visualizing political opinion formation in the web**
Authors: Michael Kaschesk, Pawel Sobkowicz, Guillaume Bouchard
Published in: dg.o '11 Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times
Pages 317-326
ACM New York, NY, USA ©2011
http://dl.acm.org/citation.cfm?id=2037556.2037607&coll=DL&dl=ACM&CFID=372997762&CFTOKEN=43551036

[3]**Opinion Mining on Social Media Data**
Authors: Po-Wei Liang, Bi-Ru Dai
Published in: Proceeding MDM '13 Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management - Volume 02
Pages 91-96
IEEE Computer Society Washington, DC, USA ©2013
http://dl.acm.org/citation.cfm?id=2509998.2510065&coll=DL&dl=ACM&CFID=372997762&CFTOKEN=43551036

Questions?